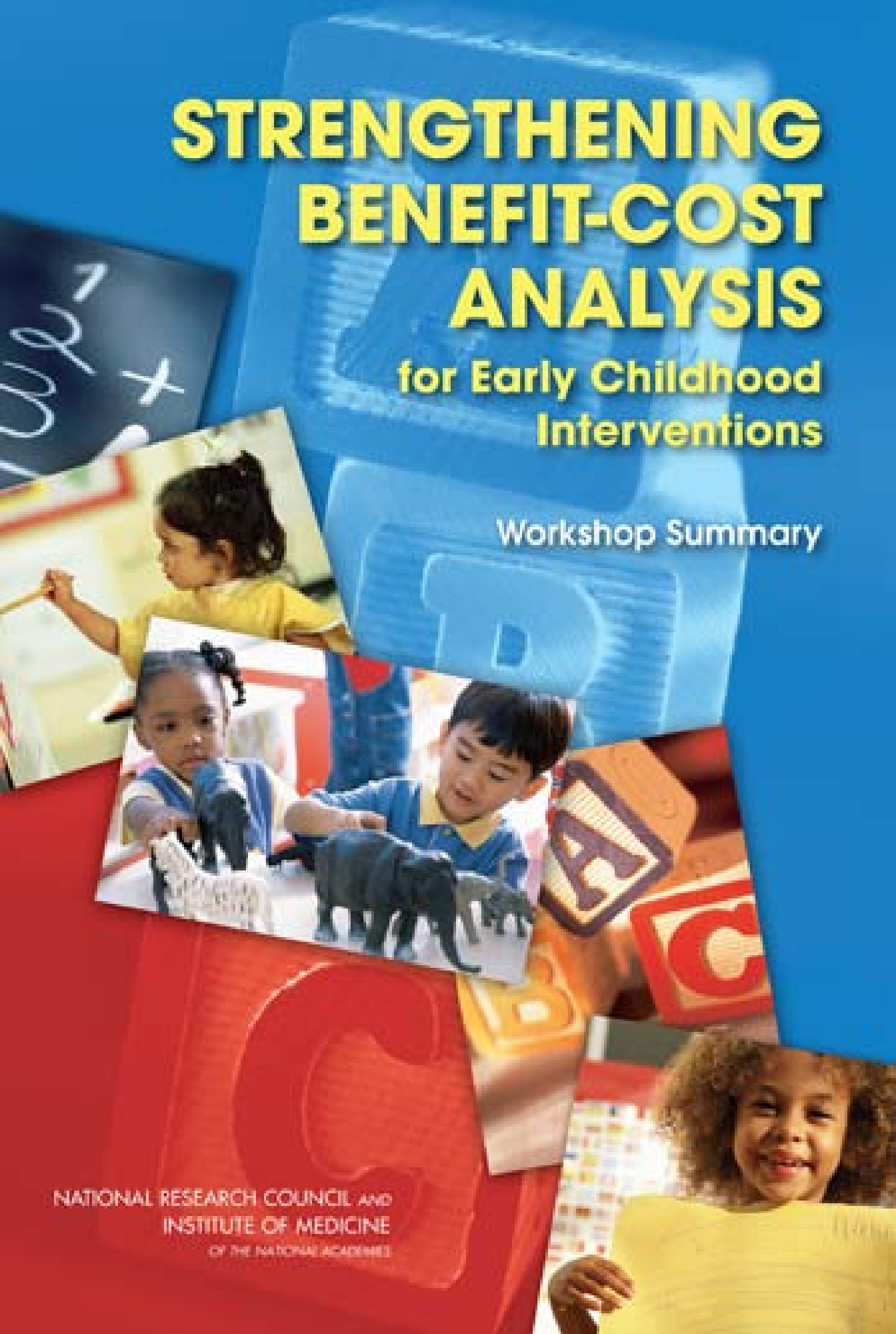
STRENGTHENING BENEFIT-COST ANALYSIS for Early Childhood Interventions

Workshop Summary



NATIONAL RESEARCH COUNCIL AND
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

STRENGTHENING BENEFIT-COST ANALYSIS

for Early Childhood Interventions

Workshop Summary

Alexandra Beatty, *Rapporteur*

Committee on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions

Board on Children, Youth, and Families

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL *AND*
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Grant No. 08-91104-000-HCD between the National Academy of Sciences and the John D. and Catherine T. MacArthur Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-14563-3
International Standard Book Number-10: 0-309-14563-5

Additional copies of this report are available from National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Research Council and Institute of Medicine. (2009). *Strengthening Benefit-Cost Analysis for Early Childhood Interventions: Workshop Summary.* A. Beatty, Rapporteur. Committee on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, Board on Children, Youth, and Families. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON STRENGTHENING BENEFIT-COST METHODOLOGY FOR THE EVALUATION OF EARLY CHILDHOOD INTERVENTIONS

Barbara L. Wolfe (*Chair*), Department of Population Health Sciences and Department of Economics, University of Wisconsin-Madison
Ron Haskins, Economic Studies Program, The Brookings Institution, Washington, DC
Robert M. Kaplan, School of Public Health, University of California, Los Angeles
Lynn A. Karoly, RAND Corporation, Arlington, VA
Henry M. Levin, Teachers College, Columbia University
Jens Ludwig, Harris School of Public Policy Studies, University of Chicago
James S. Marks, Robert Wood Johnson Foundation, Princeton, NJ
Margaret C. Simms, The Urban Institute, Washington, DC
Jane Waldfogel, Radcliffe Institute for Advanced Studies, Harvard University and Columbia University
David L. Weimer, Robert M. LaFollette School of Public Affairs, University of Wisconsin-Madison

Mary Ellen O'Connell, *Study Director*
Alexandra Beatty, *Rapporteur*
Bridget Kelly, *Program Officer*
Wendy Keenan, *Program Associate*

BOARD ON CHILDREN, YOUTH, AND FAMILIES

Bernard Guyer (*Chair*), Bloomberg School of Public Health, The Johns Hopkins University
Jane D. Brown, School of Journalism and Mass Communication, University of North Carolina at Chapel Hill
Linda Marie Burton, Sociology Department, Duke University
Angela Diaz, Department of Pediatrics and Department of Community and Preventive Medicine, Mount Sinai School of Medicine
Gary W. Evans, Department of Human Development, Cornell University
Christine C. Ferguson, School of Public Health and Health Services, The George Washington University
Sherry A. Glied, Mailman School of Public Health, Columbia University
William T. Greenough, Department of Psychology, University of Illinois at Urbana-Champaign
Ruby Hearn, Robert Wood Johnson Foundation (*emeritus*), Princeton, NJ
Michele D. Kipke, Saban Research Institute, University of Southern California and Children's Hospital, Los Angeles
Betsy Lozoff, Center for Human Growth and Development, University of Michigan
Pamela Morris, Policy Area on Family Well-Being and Children's Development, MDRC, New York
Charles A. Nelson, Laboratory of Cognitive Neuroscience, Harvard Medical School and Children's Hospital, Boston
Patricia O'Campo, University of Toronto and Centre for Research on Inner City Health, St. Michael's Hospital, Toronto
Frederick P. Rivara, Schools of Medicine and Public Health, University of Washington, and Children's Hospital and Regional Medical Center, Seattle
John R. Weisz, Judge Baker Children's Center and Harvard Medical School
Hirokazu Yoshikawa, Graduate School of Education, Harvard University
Michael Zubkoff, Department of Community and Family Medicine, Dartmouth Medical School

Rosemary Chalk, *Board Director*

Acknowledgments

This workshop summary is based on the discussion at a March 4-5, 2009, workshop convened by the Board on Children, Youth, and Families and planned by the Committee on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions. The committee members identified presenters, organized the agenda, made presentations, and facilitated discussion, although they did not participate in the writing of this report. This summary reflects their diligent efforts, the excellent presentations by other experts at the workshop, and the insightful comments of the many workshop participants. The workshop was funded by the John D. and Catherine T. MacArthur Foundation; the interest and support of Michael Stegman, director of policy, is much appreciated.

The summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We thank the following individuals for their review of this report: Steve Aos, Office of the Associate Director, Washington State Institute for Public Policy, Olympia, Washington; Matthew Neidell, Department of Health Policy and Manage-

ment, Mailman School of Public Health, Columbia University; Mariajosé Romero, National Center for Children in Poverty, Mailman School of Public Health, Columbia University; and Barbara L. Wolfe, Departments of Economics and Population Health Sciences, La Follette School of Public Affairs, University of Wisconsin-Madison.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft of the report before its release. The review of this report was overseen by Michael A. Stoto, Health Services Administration and Population Health, School of Nursing and Health Studies, Georgetown University. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author and the institution.

Contents

1

Introduction

What happens to children during infancy and early childhood has a profound influence on their experiences once they enter school and throughout life. The deficiencies that many children experience from birth to school age—in health care, nutrition, emotional support, and intellectual stimulation, for example—play a major role in academic achievement gaps that persist for years, as well as in behavior and other problems (Karoly, Kilburn, and Cannon, 2005; Kilburn and Karoly, 2008; National Research Council and Institute of Medicine, 2000, 2009). Findings from neuroscience, developmental psychology, and other fields have supported the development of many interventions designed to strengthen families, provide disadvantaged children with the critical elements of healthy development, and prevent adverse experiences that can have lasting negative effects. Early childhood interventions may focus on educational experiences in preschool classrooms, home visits, parenting education, health and wellness support, or some combination of these approaches. They may identify short- and long-term goals, such as reducing health problems, improving cognitive development and school readiness, or preventing negative behaviors like child abuse or juvenile crime.

Do these programs pay off economically? Many studies have documented benefits to children and their families, and many different models are being implemented in the United States. Policy makers are increasingly recognizing the importance of this phase of life and the potential value of early childhood interventions, and they are increasingly willing

to invest public funds in them. In a climate of economic uncertainty and tight budgets, however, hard evidence not only that such interventions provide lasting benefits for children, their families, and society, but also that the benefits translate into savings that outweigh the costs is an extremely important asset in policy discussions. Convincing analysis of benefits and costs would provide a guide to the best ways to spend scarce resources for early childhood programs. Methods for conducting the benefit-cost analysis that can provide this kind of evidence are complex in the context of early childhood, even as researchers are developing new approaches. The purpose of the workshop this report documents was to explore ways to strengthen benefit-cost analysis so it can be used to support effective policy decisions.

With the support of the John D. and Catherine T. MacArthur Foundation, the Board on Children, Youth, and Families held the workshop in March 2009 to examine strategies for strengthening the methodology for evaluating the benefits and costs of early childhood interventions. An ad hoc committee, formed to plan the workshop, was asked to explore the following questions:

- What state-of-the art examples of benefit-cost methodology can be drawn from evaluation of diverse early childhood interventions, such as home visitation programs; child care programs; Head Start; the Special Supplemental Nutrition Program for Women, Infants, and Children; Bright Beginnings; Healthy Steps; low-birthweight studies; immunization and vaccine studies; Medicaid and the State Children's Health Insurance Program; and other areas? How are benefits and costs for children identified and assessed in each program area? Are there particularly influential benefit and cost assumptions that seem important and worthy of standardizing in determining the value of selected interventions?
- How does the status of benefit-cost methodology in the field of early childhood interventions compare with studies of other vulnerable populations, such as those experiments used in assessing the impact of housing subsidies (such as Moving to Opportunities), income assistance programs (such as Temporary Assistance for Needy Families), and related activities?
- What is known about the influences of scaling up early childhood health and educational programs on both costs and benefits?
- What has been the experience with assigning a dollar (shadow) value to long-term impacts on nonmonetary outcomes like crime, health, etc.? What assumptions influence this practice and are they sensitive to specific characteristics of the populations served by selected programs?

- What lessons can be learned from the experience of other fields, such as environmental economics, to develop other approaches to program evaluation, when true benefits and costs cannot be determined within a reasonable time frame? For example, do methods such as contingent valuation analysis or estimates of "willingness to pay" offer important lessons for the assessment of the value of early childhood interventions?

This report describes the information and analysis that were presented at the workshop and the discussions that ensued. This chapter provides an overview of the nature of benefit-cost analysis and an introduction to the three primary issues: (1) evaluating early childhood interventions, (2) assessing the costs associated with them, and (3) assigning value to the benefits they yield. Chapters 2 through 6 provide a more detailed look at methodological advances and conceptual questions associated with these three basic challenges. The report closes with a discussion of the role of benefit-cost analysis in today's policy context, including how to communicate results to policy makers. Appendix A provides a glossary of technical terms used in the report, and Appendix B contains the workshop agenda and a list of participants.

AN OVERVIEW OF BENEFIT-COST ANALYSIS

There are two primary purposes for benefit-cost analysis, as committee chair Barbara Wolfe explained in her opening remarks. The first is to identify the programs or interventions that are the most effective—that is, those most likely to improve the well-being and future productivity of young children, given a particular level of expenditure. The second purpose is to guide comparisons of the benefits of investing in early childhood interventions with the benefits of other public expenditures—that is, to quantify the net long-term economic and other benefits of effective early childhood programs.

She noted that this sort of analysis is particularly challenging in the context of early childhood interventions because many of the benefits do not accrue until many years later. Numerous challenges follow from this one, such as how to identify a fair value for benefits expected far in the future, or how to extrapolate the current benefits from interventions that took place many years in the past. Moreover, it is difficult to measure many of the potential impacts of early childhood programs and to monetize their value. Nevertheless, given the Obama administration's intention to invest $10 billion per year in early childhood interventions that are evidence-based and have high benefit-to-cost ratios, she suggested, this is an ideal time to focus attention on the most up-to-date methodology

and ways to strengthen benefit-cost analysis as applied to early childhood interventions.

Although results-based accountability is a priority in any policy context, decision makers in both the private and the public sectors who are ready to invest in early childhood interventions face many additional challenges. Lynn Karoly explained that policy makers are particularly interested in analysis that can demonstrate that a dollar invested in a particular area will yield multiple dollars in savings and other benefits—that is, cost and outcomes analysis. There are four different approaches, with ascending complexity.

Simplest is a *cost analysis*, in which a program is evaluated in terms of its full economic cost (not just the direct expenditures required). A step up in complexity is a *cost-effectiveness analysis*, in which success at producing a particular outcome is part of the analysis. Cost-effectiveness analysis does not entail placing a dollar value on the outcome, however; the result is reported in terms of how much must be invested to achieve a particular outcome, as measured in the natural units for the outcome (e.g., a decrease in the percentage of youth who are held back in school or incarcerated).

The third and fourth approaches, *cost-savings* and *cost-benefit* analyses, make it possible to measure and translate multiple outcomes into dollars, which in turn make it possible to aggregate them, providing the potential to compare the total benefits of various possible investments. Cost-benefit analysis, which considers the value of an intervention both to the government (in terms of reducing the need for expensive interventions later in the life cycle, for example, reduced incarcerations for criminal behavior) and to society (in terms of broader social goods that can also have economic value, such as increases in the literacy level of a population), offers the potential for the most complete policy information—and is the most demanding type of analysis.[1]

Figure 1-1 illustrates the way benefit-cost analysis works. The two ovals on the left represent the program being analyzed and the possible alternatives to it. The analysis begins with evaluation of the resources needed to operate the program and of the outcomes or impact it may yield for children and families. The rectangular box lists some of the methods used to translate the information about resources and impact into monetary values. These methods may include calculating shadow prices for the resources used to provide the program (the implicit or true cost of continuing a program, as distinct from the monetary cost);[2] generating shadow prices in order to value any favorable (or unfavorable) effects of the program on child and family outcomes; calculating discount rates (to

[1]Benefit-cost analysis is also sometimes referred to as cost-benefit analysis.

[2]Shadow prices are discussed in greater detail in Chapter 5.

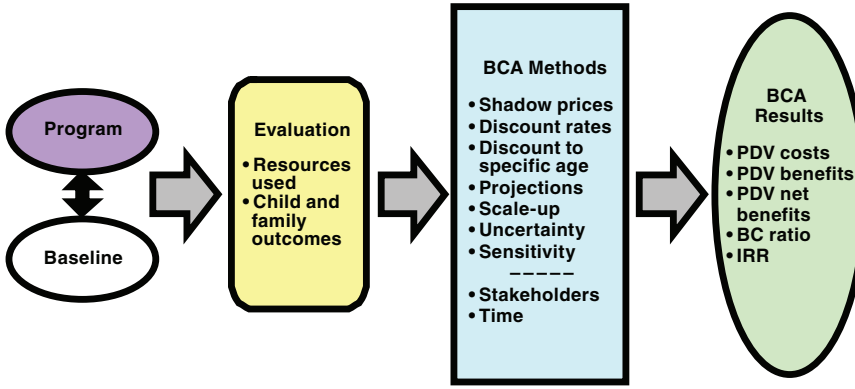


FIGURE 1-1 Elements of benefit-cost analysis.
NOTES: BCA = benefit-cost analysis, IRR = internal rate of return, PDV = present discounted value (the value today).
SOURCE: Karoly (2009).

determine the monetary value of future costs or benefits); and considering other factors, such as the impact on costs of scaling up the program. The oval at the right shows the information this analysis can yield, including the bottom line, the benefit-cost ratio.

Benefit-cost analysis might be used to look at a single program and demonstrate how that program produces a net benefit or benefits that accrue to different stakeholders. It might also be used to compare programs with one another or to compare different categories of early childhood interventions, such as programs that focus on home visits or programs that provide some form of preschool education. Similarly, the analysis might compare the effects of intervening at different stages of life, or might compare early childhood interventions with other kinds of policies or programs. As Karoly explained, however, the field is not yet at the point at which benefit-cost analyses for different early childhood interventions are comparable, nor can they be compared with those conducted for other types of social programs.

PRIMARY CHALLENGES

Some of the challenges of conducting full benefit-cost analyses are particular to the early childhood context and others are more general, but they cluster around the three primary elements of the task: (1) the need for rigorous program evaluation, (2) assessment of costs, and (3) assessment of impact. Much of the workshop focused on these three issues, and Karoly provided an overview of the questions they raise.

Need for Rigorous Program Evaluation

Evaluation is a challenge in many social policy contexts because real-world circumstances make the ideal—a true randomized controlled trial that compares two alternatives—difficult or impossible (for example, when ethical considerations make random group assignments untenable). Although a variety of quasi-experimental methods are available, these approaches are also complicated by basic questions, such as how to define the baseline against which a program or approach is to be compared. For example, if one is examining the long-term results of an early childhood program that was initiated 20 years in the past, the alternative to the intervention at that time would probably have been no program at all. Today, however, the vast majority of young children participate in some sort of preschool program, so the baseline comparison would need to reflect that. Because economic analysis depends on accurate evaluation, analysts must clearly understand the issues that complicate the evaluation.

Assessment of Costs

Assessment of the true cost of a program needs to capture not just the items shown on budget sheets, but other costs as well, such as time spent by unpaid participants or the value of lost opportunities—for example, the potential benefits of employment or other opportunities the unpaid participants could otherwise have pursued. Moreover, if approaches or programs are to be compared, the methodology for assessing the value of the necessary resources must be consistent.

Assessment of Impacts

Many different outcomes may be affected by early childhood interventions. These include changes in the children who receive the intervention (e.g., improvements in behavior or emotional or cognitive development, reduction in antisocial or risky behavior) as well as changes involving the adults in those children's lives (e.g., improved family functioning, reduction in crime or substance abuse). Most of the desired outcomes are likely to affect both children and their parents, to persist for many years, and to yield a variety of economic benefits. However, not all potential benefits may be included in benefit-cost analyses, in part because some benefits are easier to translate into dollar values than others. For example, the dollar benefits of positive emotional and cognitive development are hard to quantify. In addition, the economic value of early educational benefits doesn't emerge until later adult employment and earnings. Valuing other benefits that may not be evident until far into the future—such as lack of

involvement in crime—is also tricky. Table 1-1 shows some outcomes that have and have not been captured.

In practice, benefit-cost analysis has been conducted for relatively few early childhood interventions. In a 2005 study (Karoly, Kilburn, and Cannon, 2005), Karoly and her colleagues examined the research on 20 early childhood interventions. Of those, only seven had been the subject of benefit-cost analysis, which severely limits the field's ability to identify programs that generate returns that exceed their costs and are thus the best investments. Moreover, the analyses for the seven programs revealed that different outcomes were valued in the various studies, frequently with different methodologies. She suggested that standard practices for the conduct of these analyses would make the results much more meaningful to policy makers.

In Karoly's view, standardization would be easier for some issues, such as selecting a discount rate. Other issues—such as outcomes to be measured, the baseline to which the intervention should be compared, the length of time for follow-up analyses, and the economic values (or shadow prices) to be attached to various outcomes—are less settled. And other issues are likely to complicate standardization as well. For example, some programs begin at birth, whereas others target 3- and 4-year-olds.

TABLE 1-1 Inclusion of Benefits in Benefit-Cost Analysis

Domain	Child	Parent
Emotional and cognitive development	**Improved behavior** **Increased IQ** Increased achievement*	**More satisfactory parent-child relationship** **Better home environment**
Education	Higher promotion rates Reduced special education Increased graduation rates*	Increased educational attainment*
Work, welfare, crime	*Increased employment and income* *Lower welfare use* *Fewer arrests*	*Increased employment and income* *Lower welfare use* *Fewer arrests*
Health	Less child abuse Fewer ER visits	**Improved family planning** **Reduced substance abuse**

NOTE: Outcomes in bold are typically not captured. Outcomes with an asterisk (*) are typically used to project another outcome. Outcomes in italic are typically projected.
SOURCE: Karoly (2009).

Thus, a benefit-cost analysis for these two types of intervention would discount to different points in the life cycle, and the discounted values would not be strictly comparable, unless this difference was properly taken into account. The program design may allow for follow-up with the children who participate, their parents and siblings, or even their own offspring. Analysis that includes projected outcomes for these potential beneficiaries will be constrained by the program design, which would further complicate efforts to standardize.

Thus, the current state of benefit-cost analysis of early childhood interventions might be described as promising but still somewhat unsettled. With this overview in mind, participants considered the three primary elements of benefit-cost analysis and then considered their policy implications.

2

Evaluation

Approaches for accurately evaluating programs and drawing valid causal inferences about them are the heart of the challenge of assessing the costs and benefits of early childhood interventions. Jens Ludwig and David Deming examined two aspects of the methodological challenges of designing evaluations: (1) drawing causal inferences from studies whose designs deviate from the ideal and (2) identifying individual or group effects in analyses that include multiple outcomes and multiple groups.

MAKING CAUSAL INFERENCES

Before one can assign dollar values to a program's benefits and costs, one must first make sound estimates of those benefits and costs in the natural units with which they are normally measured, Ludwig explained. Doing so requires evidence of causal relationships, ideally collected through randomized experiments—although, as Karoly had already noted, this is not always possible. Ludwig discussed some of the methodological challenges that arise when randomized experiments deviate from the ideal design, as they often do in the real world. He also discussed alternative options for estimating causal relationships that can be used when strictly randomized experiments are not feasible.

An example of some of these real-world challenges can be observed in a recent experimental study of 383 oversubscribed Head Start centers, which began in 2002 (Puma et al., 2005; see Box 2-1 for information about

BOX 2-1
Head Start

Head Start is a federally funded school readiness program serving low-income families with young children. Created in 1965, the program focuses on preparing disadvantaged 3- and 4-year-olds for school by providing them with early education and providing their families with support in health, nutrition, and parenting. The services are supported with federal funds and delivered through locally based centers. Studies of outcomes for children who have received Head Start services show benefits that include improved performance on cognitive and academic achievement tests; increased earnings, employment, and family stability; and decreases in use of welfare and involvement with crime.

SOURCE: National Head Start Association (2009).

the Head Start Program). Ludwig noted that this study was designed to be nationally representative, with children who applied to Head Start but were not accepted serving as controls. However, as with most randomized trials, real-world complications have affected the progress of the study. Some of the participants dropped out of the study, and others may not have complied with all of its conditions—not responding to survey questions, for example. Ludwig pointed out that although the response rates in the Head Start study have been good, particularly considering that the program population is very disadvantaged, it is important to ask whether the level of attrition is sufficient to raise cautions about the causal inferences the study was designed to support.

There are several ways to approach that question. One would be to compare the baseline characteristics for the treatment and control groups. However, reassurance that they were basically similar would not provide a complete answer. Each of the baseline characteristics would have its own confidence interval (a measure of the degree of confidence one can have in the value identified, based on sample size and other factors), which suggests some uncertainty about their relative importance in explaining differences across groups and outcomes. In other words, some will be more relevant than others. To address this concern, Ludwig explained, one might use regression adjustment to examine how the estimates change when one does not control for observable baseline characteristics.

A further complication, however, is that some of the attrition in participation could be the result of factors that change over time, after the baseline characteristics are identified. A strategy for addressing that con-

cern is to consider what statisticians call worst-case bounds, which is a way of highlighting the possible effects of systematic patterns in the data that are missing on the inferences one might make from the results. To do this analysis, one first examines the best-case scenario: that all the children who received the intervention would have the best possible outcomes, and that all in the control groups would have the worst possible outcomes. By then examining the opposite assumption—that the treatment group does as poorly as possible and the control group does as well as possible—one can then see the full range of possible error.

A related concern is that not everyone selected into the treatment group will choose to participate in the Head Start Program from the beginning. The cleanest solution to that problem is to examine outcomes for everyone randomly assigned to the treatment group, regardless of whether they participate in Head Start—the "intent-to-treat" group. This approach will make it possible to identify the effect of being offered Head Start, but it will lead to an underestimate of the effects of actually participating in Head Start (because some children assigned to the treatment group do not participate). This point is often lost in policy discussions of the Head Start study results, Ludwig observed.

It is also important to be sure to compare "apples" to "apples" in comparing impacts across programs, or when doing benefit-cost analysis. In such programs as the Perry Preschool Project, for example, almost everyone assigned to the treatment group participated, so one would need to compare its effects to the effects of actually participating in Head Start (the "effects of treatment on the treated") to obtain a useful comparison. Similarly, one would not want to use the intent-to-treat effect to assess the costs of actually participating in Head Start. The fact that many children in the control group are likely to participate in some other program (rather than receive no treatment at all) further complicates the effort to carry out benefit-cost analysis.

Participation rates vary across programs and studies, and it is important to consider this point when results are compared, so that the comparisons are valid. Moreover, Ludwig emphasized, if the value of the benefits has been analyzed in terms of the entire group selected for treatment (the intent-to-treat group) but the cost estimates are based on the actual costs of enrolling a child in the program, the benefit-cost analysis will be skewed. The fact that the characteristics of those who do and do not participate may vary, in both the control and the treatment groups, also complicates the analysis.

Ludwig also explored the question of *what alternatives to randomized experiments exist in a world in which true randomization is seldom possible.* He noted that discussion of this issue often takes on an "almost theological flavor," yet it is possible in many cases to figure out the extent of selec-

tion bias in studies that are not randomized. He cited a study in which researchers compared the results of two methods. They conducted a rigorous randomized trial and then, separately, used nonexperimental data and estimation methods to evaluate the same intervention (a job training program) (LaLonde, 1986). Others have used this method in different contexts and, like LaLonde, have found substantial differences between the experimental and nonexperimental results (e.g., Dehejia and Wahba, 2002). This approach has also shown that the selection bias is likely to be application-specific—that is, it is likely to depend on the selection process for the particular program and on the quality of the data available for that program. Ludwig said that using the approach suggested by LaLonde to analyze the experimental data from the federal government's recent Head Start study would provide valuable information about the potential biases that may affect nonexperimental estimates in the early childhood education area.

One particularly promising nonexperimental approach to estimating effects is based on the idea that "nature does not make jumps," so that unusual patterns in program data are likely to indicate an effect. This approach, called regression discontinuity, has become increasingly common in the evaluation of early childhood interventions. Ludwig explained that studies using the LaLonde method suggest that *regression discontinuity* comes close to replicating the results that a true experimental study would yield. To illustrate, Ludwig used data from the early years of Head Start, which he and a colleague had analyzed (Ludwig and Miller, 2007). They noted that the counties that received early Head Start grants were those with the highest countywide poverty rates, and that no other programs at the time were addressing young children's health risks in those or other poor counties. Thus, it is reasonable to assume that, apart from Head Start, children's health outcomes should vary smoothly with respect to the baseline poverty rate. When they examined child mortality data from the period, they found that counties' child mortality rates increased along with their poverty rates up to the threshold used in awarding the Head Start grants. So the counties that received the grants saw much lower child mortality rates than the counties just above them in terms of poverty rate—an effect that clearly suggests that Head Start helped to reduce child mortality.

This approach has been used in numerous studies to estimate the effects of universal prekindergarten programs. Ludwig suggested that, for optimal results, researchers should use "experimental thinking" in designing an evaluation, even when randomization is not possible. To illustrate this type of thought experiment, Ludwig described data from a study of a pre-K program in Tulsa, Oklahoma (Gormley et al., 2005). The researchers compared the impacts for children whose dates of birth were

close but fell on either side of a cutoff for enrollment in the pre-K program. The authors assumed that the process through which the participating families choose to participate in pre-K was similar for children in both groups, but in fact the two groups of families made their decisions in two different years and thus faced different choices about alternatives. The pre-K program also may have changed from one year to the next. Thus, he explained, post facto analysis of the data yielded notably different results depending on which samples were examined. A potentially easy fix to this concern would come from collecting data on *all* children with dates of birth around the eligibility cutoff, which would be analogous to the sort of intent-to-treat analysis that is common in work with true randomized experimental designs.

Ludwig closed with the observation that program evaluation in early childhood education has come a long way since the early days of Head Start. The program began in 1965, and by 1966 the first study suggesting that it did not work was published. That study was a simple regression cross-section that compared participants and nonparticipants. In the 1990s, Currie and Thomas (1995) pushed the field forward significantly by using sibling comparisons to examine subtle differences among these groups, and a 2002 study was the first nationally representative study of the program (Garces, Thomas, and Currie, 2002). Ludwig believes that further refinements to the technology—as well as further analysis of existing Head Start data held by the U.S. Department of Health and Human Services, which workshop participants noted has been difficult to obtain—would be valuable.

EXAMINING MULTIPLE INFERENCES

Early childhood interventions are intended to have impacts in a number of domains—education, earnings, crime, and health, for example. Treatment effects also vary for subgroups that differ by gender, race, socioeconomic status, and other factors. Thus, common sense demands that studies examine multiple factors at a time, but the more factors that are included in analyses, the greater the chance for error—particularly a false positive, or Type I error. David Deming focused his presentation on *multiple inference adjustments*, which are strategies for accurately identifying individual effects in the context of an analysis that covers multiple outcomes and multiple groups.

There are two general approaches to this problem, Deming explained. The first is to create a summary index by combining various outcomes that reduces the number of tests that are part of the experiment. To illustrate, Deming used a study by Anderson (2008) of data for three prominent early childhood interventions—the Carolina Abecedarian Project,

the Perry Preschool Project, and the Early Training Project (see Box 2-2). Anderson grouped outcomes that are evident at different life stages (preteen, teen, and adult years) together, and he did the same for different categories of outcomes (e.g., employment and earnings, physical health). This way of grouping the data—by time of life—made it possible for him to standardize and compare the results across a number of studies.

The second approach, which can be combined with the first, is to adjust the probability p-values (calculations of the probability that the data indicates a significant difference) to correct for the fact that the likelihood of a false positive increases with the number of factors being analyzed. This can be done using the Bonferroni method or a method called free step-down resampling; the purpose of the latter is to account for the dependence structure in outcomes. For example, Deming explained, if two outcomes, such as high school graduation and college attendance, are most likely to be correlated but are treated as independent, some information will be lost if the probability value is not adjusted. Anderson (2008) used an approach that standardized each variable to have a mean of 0 and a standard deviation of 1 and then put them together on a single scale. The resulting effects are fairly large, Deming explained.

Making an adjustment of this sort can yield a big difference in the outcome of an analysis, but the results depend on which outcomes are covered in data collection as well as the decision about which outcomes to include in the summary index. Even more complex are the decisions about what data to include if one is attempting to use this procedure in analyzing results across several studies—so it is important to identify precise standards for selecting the studies to include, rather than simply using those that are best known, for example.

Deming closed with the point that the academic questions about the pros and cons of different statistical procedures do not always line up well with important policy questions. Different analytical approaches may yield different results, as happened, for example, with two different benefit-cost studies of the Perry Preschool Project, which produced varying results. One found larger effects for girls, and the other found larger effects for boys. But, Deming suggested, the reason for this related to the ways the two studies balanced interest in the certainty that the effects occurred at all and the magnitude of the possible effects.

His view is that while both are important, in the end, policy makers may need to be comfortable with some degree of uncertainty. If the data are adequate to demonstrate that benefits accrue from a particular program, even though it is not clear whether they will yield $100,000 or $400,000 in value, then the benefit alone may be sufficient to support proceeding with a relatively low-cost program. Even though there are many ways of clustering the data, he suggested, the potential costs and benefits

BOX 2-2
Three Early Childhood Interventions

The Carolina Abecedarian Project

The Carolina Abecedarian Project was a research study of the potential benefits of providing educational interventions to low-income children in a child care setting. The researchers identified a group of children at high risk of impaired cognitive development and randomly assigned them to receive (or not) an intensive preschool intervention between 1972 and 1977. The preschool intervention focused on social, emotional, and cognitive development, with a particular focus on language, in a full-time, year-round program. The children were followed until they reached age 21, and those enrolled in the preschool intervention showed lasting gains in IQ and mathematics and reading achievement.

SOURCES: Masse and Barnett (2002); FPG Child Development Institute (2009a).

The Perry Preschool Project

The Perry Preschool Project was a study of the effects of high-quality care and education on low-income 3- and 4-year-olds conducted by the HighScope Educational Research Foundation. This four-decade study began, in 1962, with the identification of 123 children living in poverty in Ypsilanti, Michigan, half of whom were randomly assigned to receive an intensive preschool intervention that included home visits and full-time care and education (there were intentional departures from true randomization, so this was not technically a randomized controlled trial). Followed until they reached age 40, the children demonstrated numerous lasting economic and other benefits to the treatment, including higher scores on achievement and other tests, higher high school graduation rates, higher employment rates and earnings, and lower rates of involvement in crime.

SOURCE: HighScope Educational Research Foundation (2009).

The Early Training Project

Begun in the 1960s, the Early Training Project was a study of the effects of an intervention designed to improve the educational achievement of disadvantaged children. The children were randomly selected to participate in a 10-week, part-day preschool program during the summer and to receive weekly home visits throughout the year. The random assignment evaluation showed gains for program participants in IQ, vocabulary, and reading, although some of the benefits appeared to fade over time.

SOURCE: Gray and Klaus (1970).

provide clues to the most sensible way to conduct the statistical analysis, by indicating the priority that should be assigned to various questions. The possible differences in outcomes for 3- and 4-year-olds may be outweighed by the outcomes that are evident when data for these two groups

of children are lumped together. Thus, it might be worse to fail to adopt an intervention that could significantly affect crime rates than to risk wasting a relatively small amount of money on a program that does not turn out to be effective. Participants reinforced this view. One commented that standards for Type 1 errors can be very high in studies of early childhood interventions, wondering "why are we so afraid that we might find an effect? We are giving pretty broad latitude to the possibility that there are meaningful effects that aren't passing the statistical tests."

3

Analyzing Costs

It is tempting to think that assessing the costs of an intervention is the easy part, committee member David Weimer observed, but calculating costs beyond the expenditures listed in a program budget can be difficult. Henry Levin and Clive Belfield, respectively, provided an overview of what is required and illustrated some of the issues.

RESOURCES AND COSTS OF REPLICATION

Levin indicated that studies of early childhood education rarely measure the associated costs thoroughly or accurately. First, many studies rely on budget figures, which are usually developed prior to actual expenditures, are not necessarily corrected after the fact, and rarely account for the true costs of the resources involved. Although there is fairly broad consensus among economists about how costs should be measured to obtain accurate results, that standard is seldom met in the early childhood context.

Ideally, Levin explained, there are five steps to measuring cost accurately:

1. Specify dimensions of quality.
2. Identify resource requirements to meet goals for each dimension.
3. Assess market and shadow costs for each resource.
4. Aggregate for total and obtain average and marginal cost.

5. Allocate cost burden among government support, private support, and client costs.

A comprehensive list of the aspects of the program that contribute to its quality might include the time children spend in the program (e.g., hours per day, days per week, weeks per year), the personnel ratios, the range of services supplied, facilities and materials, and so forth. Each program has its own characteristics; even when replicating a successful model, the goal is not generally to stamp out identical centers. Thus, the question of tradeoffs among certain quality features and costs arises from the beginning. Levin explained that there are various ways to document how priorities are established in the program design and replication process. This may include direct observation, in-depth interviews with the staff to ascertain what aspects of the inputs are critical, and review of program design requirements and other archival materials.

To identify the resources necessary to meet the target level of quality for selected program features, one would begin by identifying any known market prices (e.g., for staff salaries, one of the largest costs). However, current market prices may understate the long-term cost, if there is likely to be a large expansion of demand for the needed resource, or overstate it, so one must also calculate a *shadow price*. (Shadow prices are discussed in greater detail in Chapter 5.) Standard economic criteria should be used to calculate opportunity costs for participants, including those due to prospective market changes. Shadow prices also need to be calculated for any required resource for which there is not a competitive market equivalent. With all of this information, one can calculate total costs for a given enrollment goal, as well as the marginal costs that would be applicable if the program were to grow.

Sensitivity analysis—procedures to investigate the effects of various possible changes in the parameters—are important at this stage of the process, although Levin noted that this is rarely done in cost analysis. He advocated setting up confidence intervals for the cost estimates, as well as varying the quality dimensions to identify the tradeoffs and cost implications. The cost analysis will be most useful to decision makers if they can explore the cost feasibility with different budget or enrollment constraints. Levin closed with the general observation that a detailed, accurate picture of costs is just as important as a sophisticated picture of effects.

AN EXAMPLE: THE ABBOTT PRESCHOOL PROGRAM

Clive Belfield began by noting that he is often asked how much high-quality early childhood education costs. His response is that it is the wrong question—that one doesn't evaluate an investment solely in terms

BOX 3-1
The Abbott Preschool Program

A 1998 ruling of the New Jersey Supreme Court required the state to provide full-day preschool for 3- and 4-year-old children living in the 31 Abbott school districts, which are high-poverty urban districts located throughout the state (the Abbott rulings mandated numerous other educational measures as well). The court set quality standards that include qualified teachers (a state-certified teacher and an assistant in every class) and small class sizes (15 maximum); a developmentally appropriate curriculum aligned with the state's K-12 content standards; and the provision of social and health services, transportation, and support for students with limited English proficiency or disabilities. Currently there are more than 600 centers in the program, serving 38,000 children in public school or private settings or through Head Start programs.

SOURCES: Education Law Center (2007); Belfield (2009).

of how much it costs, but in terms of what is the optimal investment. He illustrated this proposition with a cost analysis of the New Jersey Abbott Preschool Program, a large-scale program in which legally mandated state standards are being implemented in a variety of settings (see Box 3-1).

The New Jersey program offered several benefits in terms of available data. In general, the program's administrative data are of high quality; line-item budget information is available for every center, as well as reports from the independent quality assurance inspections (the program uses the Early Childhood Environment Rating Scale, Revised Edition, ECERS-R;[1] see FPG Child Development Institute, 2009b). Nevertheless, some data elements are missing. For example, the program is not a full school day and does not run during the summer, but no budget data were available for the wraparound care (for the portion of the day not occupied by the program) or for the summer program. These components are funded separately, in ways that are not uniform across centers. The data do not include capital grants or parent and other nonmarket resources, nor do they include transportation costs or costs for special education services for students who need them.

Another issue is what Belfield described as contaminated resources. Funding for preschool may come from a variety of sources, including

[1]The ECERS-R system is a commercially available evaluation system that covers aspects of early childhood programs, such as physical space, program structure, routines, activities, materials, and so forth.

state allocations, child development block grant funds, welfare funds, and so forth. Many centers may be subsidized indirectly through the use of facilities or the sharing of administrative or management staff from the public education system. There are also an unknown number of younger children who are not eligible to participate for free but whose parents pay for them to participate. Furthermore, the costs for the public centers are available at the district level, whereas center-level costs are available for the private ones—but all the quality measures are at the classroom level.

Another set of problems arises with the *external validity* of the data, or means of interpreting it in the context of other programs given the lack of information on the costs of designing the program, state administration costs, or the costs to deploy and evaluate the program. Belfield also did not have information about the secondary labor costs to parents participating in the program (resulting from time away from work or changes in employment decisions, for example). Teacher pay also presents challenges—he did not have data on their benefits, pensions, or training costs. The program is growing rapidly, so short-run operating costs may not be the same as long-term marginal costs. Finally, the cost-of-education index (Taylor and Fowler, 2006), which many researchers use to compare wages and other education resources across geographic areas, was developed for K-12 education. Many factors that are different in a preschool context (e.g., large percentage of part-time workers, different credentialing requirements) limited its usefulness for Belfield's analysis of the Abbott Preschool Program.

There are several different kinds of questions one might want to answer using cost analysis, Belfield pointed out. One is to calculate the net present value, or the value of the benefits minus the value of the costs. Another is to establish links between quality and resources invested—to determine whether spending more money on specific program features will yield higher quality. Are economies of scale likely if the program is expanded? Can economies of scope be achieved, if more services are offered? Do costs vary depending on whether the programs are run through the public school system or through private providers? How do costs vary as a result of differences in local labor markets?

Four methods of cost estimating are used in litigation related to K-12 education, in which states have been charged to calculate the cost of an adequate or equitable education: (1) developing a cost function model, (2) applying data from other successful programs, (3) using an evidence-based template, and (4) asking a panel of experts to make judgments. Belfield applied three of these in his cost analysis of the Abbott Preschool Program (he deemed the professional judgment panel insufficiently reliable). He developed separate estimates for public and private centers because many of the assumptions were different for these two settings

(particularly teacher pay scales). The program in both settings operates six hours per day, nine months of the year, and Belfield noted that across the board, education costs about 20 percent more in New Jersey than in other states. His results are shown in Table 3-1.

Belfield concluded the following from his analysis:

1. Higher quality does cost more—he found that raising the quality by one unit on the ECERS scale increased costs by about 2 percent.
2. Costs were higher in private centers for several reasons. Private centers have higher facility costs, and they do not generally benefit from the cross-subsidies available to programs in the public school system. They also pay more for teachers, primarily the assistant teachers who are at the lower end of the salary scale.
3. A weak link exists between average costs and scale. Higher enrollments and multiple centers did not have much effect on costs, and costs did not vary much by district size.

He cautioned that other studies have produced somewhat different results, and in response to a question he suggested that his estimates might be capturing as little as two-thirds of the theoretical total opportunity cost of the program. Costs that are not reflected include parental resources (primarily time they spend supporting their children's experiences), costs to the state for administering and evaluating the program, transportation costs, the tax burden of raising the money to fund the program, and capital costs.

Belfield suggested several areas in need of research. More detail about possible economies of scale or scope would be useful, since expanding successful programs is a policy priority. Teacher salaries are the largest cost for preschool programs, so more understanding of the labor market

TABLE 3-1 Cost per Child of the Abbott Preschool Program for 2008-2009

	Average	Range (variation by district)
Public	$12,650	$8,920–$15,290
Private	$14,500	$11,720–$17,680
Overall average	$13,090	$7,940–$16,780

NOTE: The lower end of the overall average is lower than that for the public range because it includes some nonpublic and nonprivate centers—that is, Head Start centers that have been modified.
SOURCE: Belfield (2009).

would be useful. The burden of funding is another area that has not been thoroughly explored. For example, Belfield pointed out that a dollar of funding from one source may not have the same policy implications as a dollar of funding from another. And, he noted, ways of making cost estimating easier, such as more accurate "plug-in" numbers (estimates for particular costs that can be used to streamline cost analysis) or standardized discount rates, would encourage decision makers to take a more detailed look at costs.

4

Assessing Outcomes

Analyzing costs accurately is complex, although the established procedures for doing so apply relatively easily to the early childhood context. To assess the outcomes of early childhood interventions, however, requires careful thought about ways of measuring indirect and long-term effects. Policy makers want to base decisions about investments in early childhood programs on analysis of what can be expected in return for this investment. Advocates of these investments look for ways to demonstrate their enduring value. Ideally, accurate assessments of the potential benefits of early childhood programs would rest on common definitions of outcomes and programs and common approaches to measuring both short- and long-term outcomes. But these tools are not yet firmly in place, and researchers have been exploring a range of approaches; presenters explored their strengths and limitations and pointed to promising directions for future research.

RESEARCH QUESTIONS AND METHODS

Many studies have examined both the outcomes that are evident during or shortly after an intervention as well as the duration of these effects. W. Steven Barnett and Jeanne Brooks-Gunn described the results of several studies.

Lessons from Three Studies

Barnett described benefit-cost analyses of three of the best known early childhood programs: (1) the Perry Preschool Project, (2) the Carolina Abecedarian Project, and (3) the Chicago Child Parent Center (see Box 4-1). All three programs have been extensively studied, and Barnett presented some results from the most recent economic analyses, shown in Table 4-1, with a focus on the ways in which they approached benefit-cost analysis, their comparability, and factors that might explain their disparate results. For his summary he drew on Belfield, Nores, Barnett, and Schweinhart (2006); Barnett and Masse (2007); and Temple and Reynolds (2007). He characterized the benefit-cost ratio estimates in general terms to reflect the degree of confidence he had in them.

Barnett provided a breakdown of the value, in 2002 dollars, of the different beneficial outcomes for each of the programs, as shown in Figures 4-1, 4-2, and 4-3, and called attention to fairly large differences across the three programs. For example, the benefits in crime reduction are very large for the Perry Preschool Project; such benefits are not evident for the Carolina Abecedarian Project.

The differences in the benefit profiles reflect differences among the programs, the settings in which they operated (e.g., the baseline crime rates in the cities where the programs were located), and the populations they have served, Barnett noted. They also reflect differences in the goals of the programs, the sorts of data that were available, and the ways potential benefits were measured. Barnett suggested that researchers have made significant progress since the early 1960s, when the earliest of these

BOX 4-1
The Chicago Child Parent Center

Since 1967 the city of Chicago has provided preschool and associated support services to children and families who live in low-income neighborhoods. Eligible children ages 3-5 may participate for two years prior to entering kindergarten and may attend for half days or full days. The program addresses basic academic skills, growth and development, parenting skills, health, safety, and nutrition—parent participation in classroom activities is required. The program, which is administered by the Chicago public schools, is supported with federal funds. A federally funded longitudinal study of the program was begun in 1986.

SOURCES: For information on the longitudinal study, see Chicago Longitudinal Study (2004); for the Chicago Child Parent Center, see Chicago Public Schools (2009).

TABLE 4-1 Benefit-Cost Analyses of Three Early Childhood Interventions

	Carolina Abecedarian Project	Chicago Child Parent Center	Perry Preschool Project
Year begun	1972	1985	1962
Location	Chapel Hill, NC	Chicago, IL	Ypsilanti, MI
Sample size of study	111	1,539	123
Design	Randomly controlled trial	Matched neighborhood	Randomly controlled trial
Ages	6 weeks–5 years	Ages 3-4	Ages 3-4
Program schedule	Full day, year round	Half day, school year	Half day, school year
Cost	$70,697	$8,224	$17,599
Benefits	$176,284	$83,511	$284,086
Benefit/cost ratio	> 1	Big	Big

SOURCE: Barnett (2009).

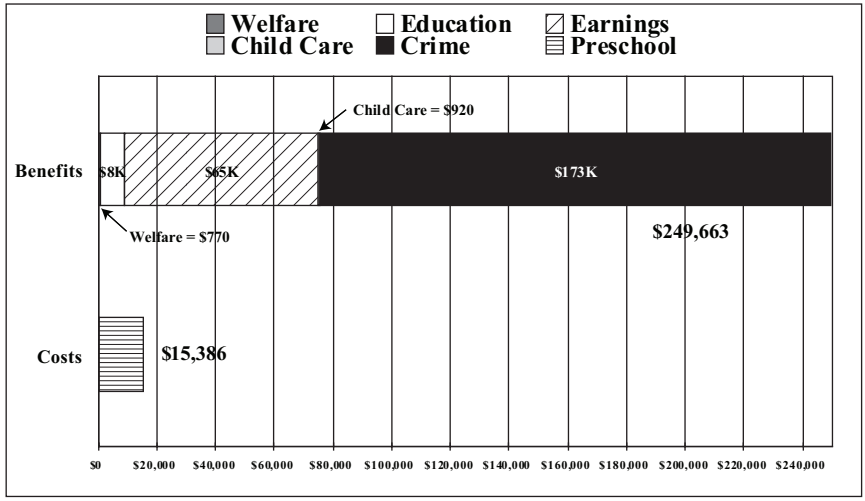


FIGURE 4-1 Perry Preschool Project: Economic return (in 2002 dollars). SOURCE: Barnett (2009).

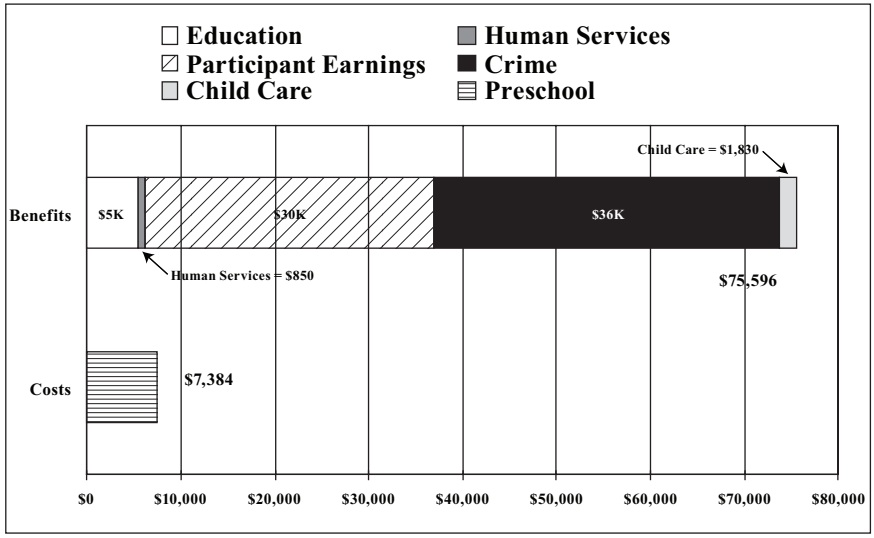


FIGURE 4-2 Chicago Child Parent Center: Economic return (in 2002 dollars). SOURCE: Barnett (2009).

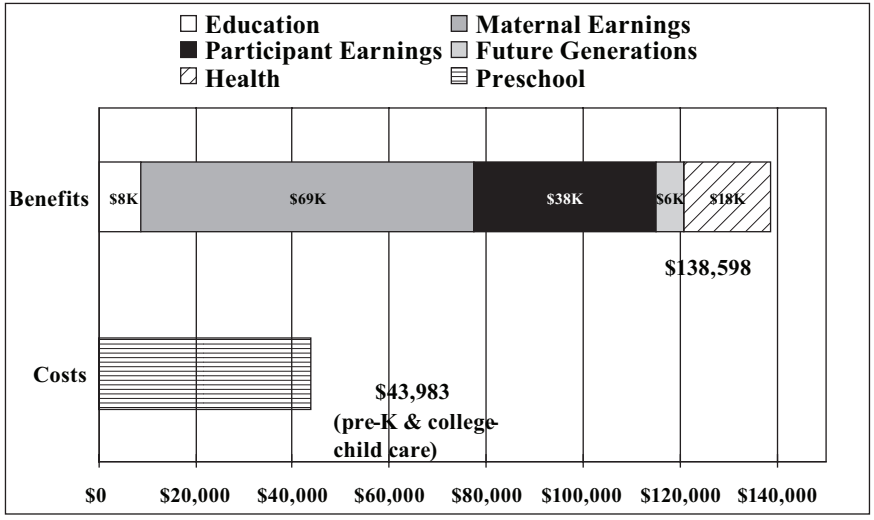


FIGURE 4-3 Carolina Abecedarian Project: Economic return (in 2002 dollars). SOURCE: Barnett (2009).

sorts of studies began, and that if they could be done over again, much more information could be gleaned.

Initial data collection for the Perry Preschool Project focused on IQ, for example, but the researchers had limited means of examining social and emotional effects (e.g., motivation, classroom behavior). They had teacher reports for the treatment group, but at the time (early 1960s) children in the control group were not enrolled in a preschool program, so no comparable reports were available for them. Similar data constraints limited the team's ability to examine many areas in which the researchers hoped to find benefits as the longitudinal investigation continued. They used proxies that seem crude today, such as using special education and grade retention costs to predict educational attainment. However, by the time the original program participants reached ages 19-40, researchers could sufficiently quantify benefits in many areas (e.g., crime reduction, welfare, educational attainment) to demonstrate a clear economic benefit. Many additional possible benefits—-such as effects on siblings or peers or improvements in family formation—could not easily be quantified.

Looking at the data for the Carolina Abecedarian Project, he noted that the initial data collection was designed by psychologists, who collected data on employment and earnings in ways that differed from economic methods. Thus, the initial data do not support analysis of the impact on maternal earnings, for example, even though the program provided full-day, year-round child care. Similarly, more could be concluded about the program's impacts on health "if we had a combination of better data and better estimates of some of the health outcomes," Barnett noted.

Looking across the three programs, Barnett had several observations. He suggested that multidisciplinary research teams—representing, for example, economics, psychology, education, and health—can ensure that the study design captures the most important information. All of the studies have very small samples, so only effects that are quite large will show up as significant, he suggested, adding that "a lot of things that are valuable get lost because of that." None of the studies looked for effects on siblings, and the measures of effects on parents are limited—again, the sample sizes are too small to support strong findings of second-order effects, but it is still possible that these are real benefits. Moreover, some direct benefits—such as increased academic success or reductions in special education referrals—are not included except indirectly, in terms of effects on earnings and reduced costs to taxpayers.

Thus, off-the-shelf estimates of value for benefits that are more difficult to quantify would make it easier to include these plug-in numbers in small-scale studies. At the same time, as programs are scaled up and large-scale analyses are feasible, it may be possible to identify small but important effects on children who are not the direct recipients of the pro-

gram (e.g., siblings, primary school classmates) and macro-scale impacts on classroom and school environments (e.g., school safety), economic growth, productivity, and so forth. A number of states are moving beyond disadvantaged children and offering programs to all children for one or two years before they enter kindergarten. These programs may have effects on the teacher labor market; working conditions; school safety, security, and maintenance costs; or even property values. In short, current arguments in favor of investments in early childhood could be made much stronger.

Focus on Improving School Readiness

Jeanne Brooks-Gunn suggested that early childhood education is important because it offers a strategy to improve outcomes for disadvantaged children. Since large numbers of disadvantaged 3- and 4-year-olds are not served by any preschool program (see Table 4-2), she said that it is important to compare outcomes for children who do or do not have access to any sort of center-based care. The biggest differences she identified were between children cared for at home and children enrolled in some kind of program. But that does not mean that quality is not important. Thus, for Brooks-Gunn, making preschool programs accessible to low-income children and ensuring their quality are the two primary goals for early childhood policy. She reviewed a range of research on outcomes that relate to school readiness to demonstrate this point, using them to highlight methodological points to consider for future research.

The effects of missing out on quality care at this age can be large, as she and colleagues found in a 2009 study in which they compared the

TABLE 4-2 State Pre-K and Head Start Enrollment as a Percentage of the Total 3- and 4-Year-Old Population

Program	3-Year-Olds	4-Year-Olds
Prekindergarten	2.7	17.3
Head Start	7.3	11.3
Special education	3.9	6.2
Other	24.8	33.6
None	61.3	31.6

SOURCE: National Institute for Early Education Research (2006).

school readiness of children in 18 cities who had been in Head Start with results for their peers who had other care arrangements (Zhai, Brooks-Gunn, and Waldfogel, 2009). The researchers used measures of attention, social competence, vocabulary, and letter-word identification, and found that children who had been enrolled in Head Start programs performed significantly better on all four measures than those in parent care or in noncenter care. Children in other pre-K programs scored as well as Head Start children on the two cognitive measures but not on the measures of attention and social competence.

Looking across several randomized trials, Brooks-Gunn has found that small-scale experiments show large effects of high-quality preschool education on school readiness (results for federally sponsored programs are somewhat smaller). The effects are evident for the children of mothers with a high school education or less but not for those whose mothers have a college degree. The effects can be larger for black children than for white or Hispanic children. Specifically, she found that if all children whose families were in poverty were in a preschool, test gaps would shrink by 2 to 12 percent for black children and by 4 to 16 percent for Hispanic children.

She also described results from the Infant Health and Development Program (IHDP), a study of interventions with low-birthweight babies that was based on the Abecedarian model. The study, which included approximately 1,000 children in 8 sites, offered children in the treatment group full-day, year-round care as well as free medical surveillance for 2 years (ages 13-36 months). Home visits and transportation were also part of the program. The study design included randomization that was stratified by birthweight, so that the researchers could compare results for children under 2,000 grams at birth and those who were heavier (but still low). Table 4-3 shows the results for both IQ and the Peabody Picture Vocabulary Test (PPVT) for the heavier children. Brooks-Gunn explained that, although the children improved in these two areas, they fared more or less like normal-weight babies in terms of health. She noted that they also saw sustained effects in mathematics achievement, reduction in aggression, and maternal employment—overall results that are greater than those for the Abecedarian and Perry Preschool projects, for example, although IHDP was only a two-year program.

Brooks-Gunn highlighted the key strengths of the study, which included faithful implementation of a tested curriculum, the collection of data on attendance (a key factor in impact), and the content of home visits. Tested curricula that are clear about the goals and activities planned and also allow for clear documentation of how they are implemented support strong analysis of effects, she explained. She noted that an independent group had developed the study design, including the randomization, the

TABLE 4-3 Infant Health and Development Program: Impacts for Children Over 2,000 Grams at Birth from Age 3 to Age 18

Age	IQ	PPVT
3 years	14.3	9.4
5 years	3.7	6.0
8 years	4.4	6.7
18 years	3.3	5.1

NOTE: All impacts were significant. IQ = intelligence quotient, PPVT = Peabody Picture Vocabulary Test.
SOURCE: Brooks-Gunn et al. (1994), McCarton et al. (1997), McCormick et al. (2006).

assessments, and the analyses, which she believes is critical to their strong findings. For example, she noted that the statistical team was firm in limiting the analysis to outcomes that were identified from the beginning of the program design.

Among the elements she would include if she were to repeat the study are measures of the quality of the care received by the children in the control groups; measures of the quality of care the treatment children received after the intervention ended, as well as the quality of their elementary education; more follow-up data (at additional developmental stages up to age 22); and data for a normal birthweight comparison group. These are needed because the outcomes depend on these factors as well as the intervention, so they should be controlled for in the analysis.

She also described some results from a study of Early Head Start that showed positive effects for children and their parents two years after the intervention ended (Chazan et al., 2007). Children showed decreased behavior problems and more positive approaches to learning, for example. Their parents showed positive effects, such as increases in reading to their children daily and use of teaching strategies, and decreases in maternal depression. Brooks-Gunn noted examples of useful data collected by the study, including detailed measures of vocabulary development, attention, and the home environment, as well as videotapes of the children interacting with their parents. She had several ideas for additional elements that would have been useful, including attendance data and more information about the curriculum.

Brooks-Gunn used these examples to highlight some of the questions the next generation of research could address:

- What differences can be attributed to differences in the setting or site in which the intervention is delivered versus differences in the population served?
- What effect does the timing or duration of the intervention have on outcomes—i.e., what is the optimal or minimal necessary amount of exposure?
- What is the optimal age to begin an intervention?
- Why are programs apparently less successful with Hispanic children and the children of immigrants?
- What elements of curriculum are important to outcomes?
- What more could be learned from studies that incorporate planned variations, in which different educational models are pursued simultaneously with comparable groups and in comparable settings, so that outcomes can be compared?[1]

ASSESSING LONG-TERM OUTCOMES

A challenge that cuts across studies and domains is identifying and measuring outcomes that persist or show up long after the intervention is completed. Katherine Magnuson and Janet Currie discussed two approaches to capturing this information.

Projecting (or Guesstimating) Long-Term Outcomes

Without a doubt, the best way to understand the long-term effects of early childhood interventions is to collect real data—that is, to follow children over time and find out what happens to them using empirical methods, Magnuson observed. But doing so takes time and money; therefore, it is useful to explore other ways of estimating long-run outcomes. Complex procedures are involved in developing such estimates for complicated production functions. Inputs at different ages, and of different sorts and magnitudes, may have differential effects on health, cognition, language, and behavior. Most early interventions explicitly or implicitly target more than one domain, or they might be expected to have effects that spill over from one domain to another. For example, in an effort to improve cognitive functioning and academic achievement, a program might teach children to focus and concentrate, which would be likely to produce other benefits as well.

Several methods exist to resolve this complexity, and all yield at best rough approximations. One approach, used by Krueger (2003), attempted to estimate the later earnings benefits of reducing class size. Krueger

[1]A planned variation study of Head Start programs is described in Kennedy (1978).

looked at studies that linked early achievement to later earnings and applied the percentage (8 percent) to data from the Tennessee STAR (Student Teacher Achievement Ratio) experiment on class size. This approach could be adapted to produce rough estimates for other predictors and outcomes, Magnuson explained, but there are a few complications in applying it to early childhood interventions.

One question is whether outcomes for an intervention in early childhood are different from the outcomes of the same intervention with older children. For example, the behavior issues of 2- or 3-year-olds, 4- or 5-year-olds, or 8-year-olds are likely to be different and to decrease over time. Thus, it is important to consider children's developmental progressions in measuring effects on behavior. A more fundamental problem with this approach is the lack of sources of nationally representative, high-quality data on early childhood achievement, behavior, attention skills, and other elements, together with wage data for later years, which are needed for this type of analysis.

Adapting this approach in a two-step analysis could provide an answer to some of these concerns, Magnuson suggested. Here, one would first link an early childhood outcome, such as achievement at age 5, to a more proximate outcome, such as adolescent achievement or high school graduation. The latter outcome could then be linked to an outcome of interest, such as adult earnings.

The advantage of this approach is that more data are available to establish the magnitude of the two links, although Magnuson acknowledged that a variety of measurement issues contribute uncertainty at each step of the process. For example, which measures and samples provide the most accurate results was unclear and open to discussion. Another point that needs consideration is which research designs best approximate the causal effects, because arriving at good estimates depends on accurately identifying the magnitude of causal links. Put another way, the results are only as good as the studies from which the data are drawn. Finally, the model can map only effect pathways that have already been measured—overlooking other possible pathways that link early childhood experiences to later outcomes. Nevertheless, the two-step method is flexible enough to be adapted to examine a variety of outcomes, and it provides a transparent logic model for explaining how the effects work.

Another way to develop estimates is to leverage experimental evaluations from studies of other programs that have examined long-term outcomes. Magnuson used data for the Perry Preschool Project to illustrate how this can be done. The operating assumption is that the effects are likely to be proportional. So, using data on the Perry Preschool's effects on measures of early achievement of language and on later earnings, one can calculate the probable effects of other programs for which only early

data are available. The Perry Preschool's effect on the PPVT was .91 standard deviation and on lifetime earnings was $59,000 (in 2006 dollars); one can use program impacts on the PPVT from another program and calculate a probable (proportional) effect on earnings. This model, Magnuson explained, has the advantage of not requiring that all mediating pathways to the long-run outcome be modeled, so it doesn't require assumptions about which pathways explain the effects. However, the validity of proportional relationships has not been empirically tested, so it is a large assumption to make. Moreover, the ways in which the benchmark program results in long-run outcomes, and the population for which it was studied, may have unique characteristics that account for its effects.

Table 4-4 shows the results Magnuson calculated using each of these methods, including the two-step version using two different intermediary measures—adolescent achievement skills and high school completion. She suggested that all are reasonable methods for obtaining a rough esti-

TABLE 4-4 Comparing Approaches

Program Impact in Early Years	A1[1] (Krueger)	A2[2] (2-step Ach)	A3[3] (2-step HS)	A4[4] (Prop. to Perry)
	PV Earnings in 2006 Dollars			
1 SD reading	$40,330	$20,160	$9,720	$64,835
.5 SD reading	$20,160	$10,080	$4,862	$32,417
.2 SD reading	$8,070	$4,030	$1,945	$12,967
	Fraction of PV of Lifetime Earnings			
1 SD reading	.08	.04	.02	.09
.5 SD reading	.04	.02	.01	.04
.2 SD reading	.02	.008	.004	.02

NOTES: Present value of lifetime earnings ($508,104) is calculated for a sample that is 50 percent high school graduates and 50 percent high school dropouts. All columns present 2006 dollars with 3 percent discounting to age 5; columns 1-3 assume 1 percent wage growth. PV = present value, SD = standard deviation.

[1]A1 represents a variation on Kreuger's (2003) method.

[2]A2 uses a two-step approach with adolescent achievement skills as the intermediary outcome.

[3]A3 uses a two-step approach with high school completion as the intermediary outcome.

[4]A4 assumes that effects will be proportional to those found in Perry Preschool.

SOURCE: Magnuson (2009).

mate, although each has strengths and limitations. They yield different results because each entails making a variety of assumptions and thus reflects the pathways the analyst views as important and outcomes he or she expects to see.

Leveraging Administrative Data

Janet Currie also addressed the problems resulting from the lack of longitudinal data: that they are expensive to collect, that attrition of participants over the years can be a serious problem, and that, by definition, the data produce answers only years after the intervention begins. She offered three approaches to make better use of existing data: (1) posing new questions that can be answered using existing data, (2) merging new information into existing data sets, and (3) merging several existing data sets. She noted that in the United States it can be difficult to obtain the relevant administrative data for these kinds of analyses, but that these approaches have become increasingly common in other countries—particularly Canada and the Scandinavian countries.

Using two studies as examples, Currie discussed the pros and cons of the first approach. Garces, Thomas, and Currie (2002) asked whether a group of adults for whom they had data from the Panel Study of Income Dynamics (PSID) had ever been enrolled in a Head Start program or had attended another preschool, while Smith (2007) compared their health status in earlier years. The PSID was useful for this purpose because it is a long-running study that provides rich information, including a large sample and data from siblings. Currie also observed that retrospective data may contain errors, but that there are strategies to address that problem. For example, one can compare reported participation rates or distributions of characteristics to available confirmed records. She also noted that it is possible to examine only outcomes that are already reported—that is, one cannot go back and examine some other factor, such as family life, for which no data had been collected.

Another study demonstrates the potential of merging new data with existing data sets, which is typically done by geographic area. As discussed in Chapter 2, Ludwig and Miller (2007) used data from the National Education Longitudinal Study of 1988 (NELS) to study the effects of Head Start. The 300 poorest counties in the nation received assistance in applying for Head Start funds when the program was initially rolled out, so they were more likely to have Head Start programs than were slightly richer counties. By drawing on vital statistics and census data, the researchers were able to establish that counties with Head Start programs had lower childhood mortality rates and higher education levels than did poor counties without the program.

The third approach—merging administrative databases—requires the use of confidential information (data with personal identifiers). If this obstacle can be overcome, this approach can provide valuable information. Currie and colleagues (2008) merged data from Canadian public health insurance records with data from the welfare and education systems to examine possible links between health problems in early childhood and future welfare use or lower educational attainment. They found that major health problems at ages 0-3 are predictive of both poorer educational attainment and welfare use, primarily because poor health at early ages is predictive of poor health in the later years. They also found that mental health problems were much more predictive of future welfare use and lower educational attainments than physical health problems.

This approach allowed the researchers to work with a large sample and to create objective indicators—the data were recorded by medical providers. The approach allows sibling comparisons and long follow-up periods. However, these data sets do not provide much background information. The health measures were dependent on whether or not individuals sought care for a particular problem, although, in this Canadian sample, virtually all children received health care. And, of course, this approach can be used only if administrative data can be accessed by researchers.[2]

Currie pointed out that privacy concerns are making it more difficult to obtain administrative data, just as methods for using them for new purposes are becoming more feasible. For example, natality data used to include county of birth, but since 2005, this has not been the case. She suggested that creators of large data sets should be sensitive to the fact that their data may well be used to answer questions that have not yet been considered. Thus, they should retain information that can make linkage after the fact easier. For example, geographic identifiers (census tract or zip code) should be retained. Participants could be asked to sign informed consent forms even if they are not immediately needed, since they generally cannot be obtained retrospectively. She also advocated further research on methods for making sensitive data available without compromising people's privacy. Data in small cells—perhaps for rare outcomes—could be suppressed, for example, or a small amount of statistical "noise" could be added to public-use files to obscure identifications. Data use agreements, such as those used in the National Longitudinal Survey of Youth (NLSY) or NELS, can allow researchers access as long as they agree to various restrictions, such as signing data use agreements, or using only a standalone computer (not a network) for the analysis. Data

[2]She also cited Black, Devereux, and Salvanes (2007) and Doyle (2008) as examples of studies that use the merging of databases.

swapping—in which those who hold confidential data run a specific analysis for other researchers and then strip out identifying information—is another approach.

In Currie's view, a great deal of valuable information is locked up in administrative data sets that are not currently accessible—and making use of them could be a cost-effective way to answer important questions. Many participants supported the idea, noting, for example, that "we are not going to be reproducing the Perry Preschool study any time soon, and we don't want to wait around for 40 years [but] we are going to be implementing these programs."

Looking at the back-of-the-envelope estimates Magnuson had described as well as Currie's linkage approach, a participant noted that they are "useful—if you know what the cost is. If even a rough estimate that you think is an underestimate is still higher than the cost of the program that you are thinking about," you have enough information to go forward. Moreover, these kinds of approaches make it possible to look at far larger samples: "We can break it down for different types of children so we can look at whether there are differences in these patterns by children with different backgrounds or different ethnicities—true data may be best, but we are never going to have large enough samples given the cost of collecting it."

5

A Closer Look at the Problem of Valuation

The questions that arise in assessing benefits and costs for early childhood interventions have emerged in other contexts, and the workshop was designed to consider relevant insights and examples. David Weimer provided a detailed look at the development of shadow prices in general. Myrick Freeman, Philip Cook, and Donald Kenkel discussed the ways monetary values are assigned to outcomes in three sectors, respectively: (1) environmental economics, (2) criminal justice, and (3) health.

THE IMPORTANCE OF SHADOW PRICES

Shadow prices are a means of (1) converting projected program impacts into social benefits (which can be measured in terms of society's willingness to pay for them) and (2) converting program resources into social costs (measured as opportunity costs). Many plausible, but imperfect, shadow prices are available in the early childhood context, mostly based on data from long-term experiments, such as the Perry Preschool Project. These are extremely useful, Weimer suggested, but they cannot be "the answer to all our problems because we're just never going to have enough resources to do enough of them." Moreover, he stressed, studies of a single program by definition can answer only a constrained set of questions, from the point of view of decision makers. More "wholesale" experiments are necessary to provide the basis for useful shadow prices—which are key to benefit-cost analysis.

In general, different kinds of information can be used to calculate shadow prices. One is the market price of various resources in the early childhood context, such as wages and benefits for teachers. If the market is distorted, as, for example, when a preschool program does not pay market price for the use of school buildings, an adjustment might be made by calculating opportunity costs. Economists might also use indirect methods to calculate values for which there are no clear market prices (missing markets). For example, they might calculate the statistical value of a life year, or infer the *contingent value*—the amount people say (usually in response to survey questions) they would be willing to pay for a resource that does not have a market value. In the context of early childhood, that might mean calculating opportunity costs for volunteer time or benefits of improved educational outcomes or a reduction in crime.

Theoretically, the ideal way to conduct a benefit-cost analysis for early childhood interventions would be to use a long-term random assignment experiment, much like the Perry Preschool Project, "except bigger and perhaps more geographically representative," Weimer explained. These data would make it possible to predict the impact of other similar programs, using shadow prices to estimate earnings changes, quality of life changes, willingness to pay for various benefits, and so forth. However, Weimer suggested that this model is not actually ideal from a public policy perspective. Long-term studies are expensive, so sample sizes tend to be small, and such studies are relatively rare. Researchers often encounter problems with attrition and have difficulty accurately taking into account long-term shifts in the context—such as changes in the sorts of alternatives that are available to the program being studied. And, of course, results are often delayed.

Weimer offered several alternative approaches. First, work could be done to develop better shadow prices for the early childhood context. He pointed out, for example, that the RAND Health Insurance Experiment (Manning et al., 1987) provided a way of developing estimates of the price elasticity of demand for health care. Existing early childhood studies provide observational data that could be used in a similar fashion to link program effects to outcomes, such as school completion, for which shadow prices may be more readily available. More work is also needed in the development of shadow prices for willingness to pay for societal benefits, such as reducing poverty, using contingent valuation techniques.

Another promising approach is to improve strategies for linking observable outcomes to a wider array of social benefits (Weimer and Vining, 2009). Decades ago, Haveman and Wolfe (1984) used a household utility approach to estimate the nonlabor market benefits of schooling (such as reduction in crime, efficiency of consumption). They calculated a monetary value for such outcomes as children's cognitive develop-

ment, the use of contraceptives, consumption efficiency, and improvements in health. They concluded that the nonlabor market gains were approximately equal to the labor market gains, and their rough estimate suggested that each additional dollar in earnings resulting from an intervention produces about an additional dollar in social benefits. In another study, the same researchers looked at an even wider array of effects that schooling might have, including effects on the schooling of participants' children, on family members' health, on participants' daughters' fertility, among others (Wolfe and Haveman, 2001). However, few others have attempted this sort of analysis.

A third approach is to improve ways of linking immediate impacts to future benefits. By drawing on other empirical research, Weimer explained, one could link short-term impacts, such as school readiness, with longer term outcomes for which shadow prices are available, such as school completion. The next step (Weimer acknowledged that Katherine Magnuson had suggested a similar analysis) would be to compress the chain of causality to produce a shadow price for the immediate impact that can be used in comparing alternative programs. Another example of this approach is a meta-analysis conducted by Aos, Miller, and Drake (2006), using studies of the criminal justice system.

For Weimer, the bottom line in a policy context is to obtain the best estimates possible to support decision making. Doing that requires less emphasis on whether a program worked in the past and more on mining its results for indications of what will work now and in the future. It is also important to decrease the cost of conducting benefit-cost analyses, because, until that happens, "we're not going to have enough of it." The shadow prices are key to efficient, low-cost analyses, but, he noted, "we have to go outside of our discipline sometimes to do that."

EXAMPLES FROM OTHER SECTORS

Environmental economics, criminal justice, and health economics are three fields that have made considerable progress in the use of benefit-cost analysis, and each offers insights that could be useful in the context of early childhood.

Environmental Economics

The degradation of environmental resources—such as clean air and water, biodiversity, a healthy ecosystem—was an early impetus for economists to develop ways of assigning monetary value to benefits or resources that are not traded for money. Myrick Freeman described the two primary methods of using nonmarket valuation to assess the effects

of environmental policies in situations in which no market prices are available for analysis. The objective for these methods, he explained, is to estimate the willingness of people affected by a government policy to pay for the benefits it is expected to yield.

One set of methods, the *revealed preference* methods, examines the choices made by people who are or may be affected by a policy. Specifically, these methods use data about people's choices to identify the implicit prices that they would pay to achieve a particular outcome. These methods are based on the rational choice economic model[1]—that is, analysis based on the assumption that when people are rational and have enough information to make an informed choice, their marginal willingness to pay for an environmental improvement indicates the economic value they attach to that improvement. Thus, if the marginal implicit price can be estimated, the marginal willingness to pay can be inferred. One example of how this works would be to analyze avoidance behaviors: People's willingness to pay to avoid the risk of waterborne disease can be inferred from the prices they pay for water filters or for bottled water. Similarly, one might examine the relationship between the risk of death or injury on the job and wages to identify people's marginal willingness to pay to reduce these risks. This is done using hedonic wage models; similar models are used to examine housing prices to identify people's willingness to pay to live near such amenities as a park, a waterfront, or a school.

The other set of methods is the *stated preference* method, in which people are asked hypothetical questions about their preferences and willingness to pay for various benefits. (Contingent valuation, discussed above, is one form of this approach.) This can be done in various ways. One could provide a reasonably detailed description of the resource or benefit in question, ask people to supply a dollar figure or choose from options, and then calculate the mean response. One could ask whether respondents would or would not be willing to pay a particular amount, perhaps following up with a second amount, depending on the answer. There are also various ways of asking people to rate or rank a set of alternatives and, using discrete choice models (mathematical functions that take into account the attractiveness of various options), to predict the tradeoffs people are willing to make between price and the selected attributes.

However, Freeman explained, the stated preference models are all controversial, particularly in the context of litigation regarding assessments of damage to natural resources. The lawsuit that followed the

[1]The workshop did not provide an opportunity to examine other economic assumptions, such as those suggested by behavioral economics, in which psychological and ideological motivations are explicitly considered.

Exxon Valdez oil spill in 1989, for example, generated considerable dispute and research to advance the techniques. Observers have questioned the reliability of the responses, which is difficult to assess because true values are not available (i.e., hypothetical choices may not accurately represent the choices people would make if they faced a real-life decision). The possibility that respondents have an incentive to misrepresent their values in some strategic way has also been raised, as have questions about whether respondents can be assumed to have full information about the alternatives (the latter question would be relevant to the revealed preference approach as well).

An additional problem is how to address the likelihood that respondents do not always have well-defined preferences regarding the options they are asked about and may make them up on the spot. Such decisions are likely to be influenced by the information they are presented with. Studies of the issue find a degree of consistency in people's responses, suggesting that they tend to have formed preferences that guide their responses. While researchers have developed various strategies for addressing these concerns, Freeman was clear that the results are definitely affected by the way questions are framed, and that it is both "easy to do a bad study and very hard to a good study."

To illustrate the application of some of these methods, Freeman described three analyses of the benefits of reducing childhood exposure to lead, two of which were conducted by the U.S. Environmental Protection Agency. Studies dating back to 1985 have established clear health benefits for both adults (reducing hypertension and risk of heart attack) and children (avoiding cognitive deficit and other health problems) from limiting lead exposure. These health benefits have economic benefits, such as reduced education and medical costs, improved lifetime earnings, and reduction in antisocial behavior. (Freeman cited Schwartz et al., 1985, and U.S. Environmental Protection Agency, 1997.) Similar studies have examined the value of reducing childhood exposure to mercury.

These examples highlight several issues regarding the economic valuation of early childhood interventions. The first is the question of choosing a normative perspective, as reflected in whose willingness to pay should be counted—the child's, the parents', or that of a hypothetical child endowed with the financial resources and cognitive abilities of an adult. Different analytic methods, Freeman explained, imply different normative perspectives, so it is important that researchers consider this point and make its implications clear in their analysis.

A second issue is the challenge of capturing third-party effects. A potential crime victim can be presumed to have some willingness to pay for the reduction in crime that could result from reductions in lead exposure to children, for example, but the potential victims cannot be identi-

fied or asked in advance. And third, there might also be societal benefits from early childhood interventions that people may be willing to pay for, even though they will not personally be affected by the intervention. For example, would people in general have a willingness to pay to see a reduction in the prevalence of childhood obesity? While these issues remain somewhat unsettled, researchers and policy makers have made good progress in valuation in environmental economics that could be a useful resource for research and policy work in early childhood.

Criminal Justice

Attaching value to the impacts of crime is done using similar methods and raises many of the same points, Philip Cook explained. He illustrated valuation for crime-related questions with a simple example. If a residential community is considering hiring a guard in order to reduce crime rates, their decision would be based on whether the value of having the guard is greater than the cost (the guard's salary). Looking first at property crime, the residents could begin with the property value that might be lost to theft or vandalism without the guard. Even in this simple case, though, there are complications, such as whether or not the residents have property insurance (and whether the rates might be affected by the presence of the guard), how risk averse they happen to be, the possible sentimental value of items (apart from their monetary value), and other negatives associated with crime, such as invasion of privacy or the inconvenience of replacing lost items.

In thinking about violent crime, the residents would again begin with direct costs that could be averted, such as medical care and lost earnings. Indirect costs, such as the fear and disutility the residents anticipate from being a victim, are likely to be significantly higher but difficult to monetize. If all the residents are willing to pay the cost of hiring the guard, then it is clear that the benefit exceeds the cost, but this answer is imprecise because individual residents might be willing to pay very different amounts. A market test of the decision is whether the value of the property in the community goes up in spite of the fee for the guard; this result would be evidence that a larger group of people (beyond the residents) believes the community is more attractive with the guard.

But it is not clear that this evidence of the community's collective willingness to pay for the guard is the same as the value they attach to the social benefit of reducing crime, Cook explained. Looking more broadly makes other issues apparent. Hiring a guard for one community might simply displace the crime to other communities, so the net benefit to society at large would be zero. Other factors, such as people's views about the

other effects having a guard might have on the community, would also be factors in the valuation.

Having introduced some of the primary issues, Cook described a study in which he and Jens Ludwig examined people's views of policies designed to reduce gun violence (Cook and Ludwig, 2000). Using the stated preference model, they asked respondents how they would vote on a policy that was described as having the potential to reduce gun violence by 30 percent. Using randomized samples, they told respondents the cost would be a tax increase of $50, $100, or $200 and used follow-up questions to refine the responses. With these data they were able to trace a demand curve and calculated that the 30 percent reduction in gun violence was worth an average of $240 per household, or approximately $1 million per shooting. In a similar study Cohen and colleagues (2004) found that preventing a burglary was worth $25,000, preventing a robbery was worth $232,000, preventing a rape was worth $237,000, and preventing a murder was worth $9,700,000.

Some have questioned these numbers, but potential benefits are nevertheless very large. Picking up the example of reducing children's exposure to lead, Cook suggested that if a generation grows up with a lower average criminal propensity because of widely decreased lead exposure, the result will be "not just less crime but an array of outcomes that will be interacting with each other: less crime, lower response costs from the public and private sectors, and so on." Valuing these system-wide benefits is challenging but nevertheless important.

For Cook, "the bottom line is that there is no very reliable approach in this area—it is tough to get stable numbers that can be reproduced." Looking forward, he suggested that continued development of analyses of willingness to pay will be important, as will further developing analogies between crime and disease.

Health Economics

In research on the economics of health, Don Kenkel explained, two methods are used for valuation: (1) cost-benefit analysis based on willingness to pay (using shadow prices as described above), and (2) cost-effectiveness analysis. The first, cost-benefit analysis, is done the same way whether the context is health, environmental policy, or criminal justice, but in practice it is less common in the health context, so Kenkel focused on cost-effectiveness analysis.[2] The simplest version of cost-effectiveness

[2]Kenkel cited Tolley, Kenkel, and Fabian (1994) as a source for the cost-benefit approach in health. For more information on the second approach, he suggested Institute of Medicine (2006).

analysis is to relate the costs of an intervention to a direct effect, such as the cost per cases of cancer detected using a particular screening method. If one takes the extra step of considering the utility of particular outcomes, or people's preferences, it is called cost-utility analysis.

In considering utility, economists may use a health-adjusted life year—a way of measuring both the quality and the length of lives saved by a health intervention—as a common unit to represent the value of health. This tool makes it possible to consider not only how many lives were saved, but also whether it was the life of an 80-year-old or a 20-year-old, and whether the remainder of the life was spent in bad health or disability or in good health. (He noted in response to questions that adjusting for age also comes up in cost-benefit analysis based on monetary willingness to pay for health and safety. Making monetary adjustments based on age is complex; not only is there a range of views about whether it makes sense to adjust for age, but also there is no consensus on how best to do it.)

A commonly used example of the health-adjusted life year is known as the quality-adjusted life year (QALY). Stated preferences are used to construct this tool. Here, however, respondents are not asked about their willingness to pay for an outcome but are presented with a "standards gamble" as a way of finding out how they would weigh the risks and benefits of staying in a suboptimal state of health or risking a worse outcome in pursuit of an improvement. A stark example would be to ask respondents whether they would risk an operation that could restore them to perfect health but carries a 10 percent risk of death. Another approach is to ask respondents to report the relative value they would place on, say, 10 years lived in suboptimal health versus 1 year in optimal health. From the responses, researchers can estimate the relative value of different outcomes.

Kenkel described an early example of cost-utility analysis of a medical issue—childhood lead poisoning—conducted by Glotzer, Freedburg, and Bauchner (1995). The researchers examined the cost-effectiveness of several different approaches both to testing children for exposure and to treating those who are exposed, including remedial education to address cognitive disability and chelation to remove the lead (a painful and expensive procedure). They estimated the value of detecting and medically treating lead poisoning at approximately $1,300 per QALY gained. When they factored in the cost of remedial education that would not be needed if the poisoning was prevented, they found that the intervention was cost saving.[3] A study of the cost utility of screening for fetal alcohol spectrum

[3]Kenkel noted that the study does not reflect later findings that questioned the effectiveness of chelation.

disorder provides another example (Hopkins et al., 2008), in which the researchers found an incremental cost-effectiveness ratio of about $66,000 per QALY.

How were these values for a QALY calculated? In analyzing the effects of lead exposure, Glotzer, Freedburg, and Bauchner (1995) assumed that life with a lead-based disability would be counted as 77 percent as valuable as life without—and thus calculated a QALY weight of 0.77. However, they based that figure on what Kenkel described as "thin evidence," a survey of 13 pediatricians and pediatric educators at their own institution. The QALY weight calculated for fetal alcohol spectrum disorder was 0.47, in this case based on a survey of 126 children and families about their experiences with moderate to severe dysfunction resulting from the disorder. Kenkel noted that there are many other studies using QALYs that may have stronger evidence to support these sorts of calculations, but that applying the approach to children is not a well-developed procedure.

Although some aspects of the approach are not fully settled, cost-utility analysis is widely used in health and medicine. Kenkel noted a registry housed at Tufts University, the National Institute on Clinical Excellence in Great Britain, and efforts in other countries to collect this kind of evidence of cost-effectiveness for pharmaceuticals and other medical options. The U.S. Office of Management and Budget provides guidance for using this approach in regulatory analysis (Executive Order 12866), and Kenkel suggested that all federal agencies should prepare such an analysis as part of any rulemaking related to public health and safety. The Institute of Medicine, he noted, has also made recommendations for using measures of cost-effectiveness to support federal regulations (Institute of Medicine, 2006).

Nevertheless, Kenkel noted, some are skeptical about benefit-cost analysis for health issues. The analysis is a way of getting around putting an explicit monetary value on health effects, but some question whether estimates of willingness to pay are reliable. Willingness to pay may vary with income, and many are uncomfortable with the idea of connecting the allocation of health care to personal income. Still, the analysis does provide implicit monetary values that are useful. The challenges of calculating willingness to pay in the context of questions related to morbidity and mortality are not conceptually different from the challenges of valuation in other contexts. Some would also argue, Kenkel suggested, that questions about health are special and should not be subject to utility-based analysis. However, cost-effectiveness, or cost-utility analysis solves this problem because its purpose is to identify the optimal way to allocate limited resources—to produce the maximum health benefits for a fixed amount of money.

Kenkel observed that "theoretical purity doesn't necessarily translate

into effective persuasive policy advice. . . . It is very clear that estimates of cost savings from interventions have a lot of persuasive appeal." He closed with a quotation from a 1971 paper called Evaluation of Life and Limb: A Theoretical Approach, "In view of the existing quantomania, one may be forgiven for asserting that there is more to be said for rough estimates of the precise concept than precise estimates of economically irrelevant concepts" (Mishan, 1971).

6

Generalizability of Benefit-Cost Analyses

Although methods for estimating costs and valuing outcomes raise many important conceptual issues, they are of less interest to policy makers than accurate general conclusions that can be drawn from a body of research. As the methodological discussions suggest, generalizing from studies of the benefits and costs of early childhood interventions poses its own complexities. Mark Lipsey discussed the potential value of meta-analysis for this purpose, and Howard Bloom examined some broad design and analysis considerations.

META-ANALYSIS

Lipsey began by suggesting that, when research findings can be generalized, it means that the same intervention will produce the same or nearly the same effect despite variation on some dimensions, such as the characteristics of the providers or recipients of the intervention, the setting, and perhaps certain nonessential features of the intervention itself. Ideally, a generalization is based on a relatively large sample of research studies of effectiveness; the studies will have used representative probability samples of the population of interest and random assignment in order to provide both internal and external validity in the same study. That ideal is hard to attain, Lipsey noted. Studies that have evidence of internal but not external validity and that use samples that are "not-too-unrepresentative" are more common.

Meta-analysis across such studies provides a next-best approach to

drawing as much information as possible from multiple studies, particularly when the interventions are similar and the samples represented are diverse across the studies. The most important feature of a meta-analysis is representation of the effects on a certain outcome in terms of a standardized effect size that can be compared across studies. Analysis then focuses on the distribution of effect sizes—the central tendency of that distribution and also the variation around that mean. The key question is the extent to which that variation is associated with or can be explained by moderator variables, such as differences in setting, subject characteristics, and so on. So, in essence, meta-analysis is the empirical study of the generalizability of intervention effects.

A few issues make this analysis challenging. First, the question of what constitutes the "same" intervention is complicated. Few interventions are crisply and unambiguously defined, and the developers of an intervention may modify it as they learn from experience. At what point are the modifications sufficient to produce a different intervention? In general, meta-analyses are designed to focus on a *type* of intervention defined generically, rather than in terms of a specific intervention protocol. However, there are no formal typologies to which researchers could turn for grouping similar interventions in areas like early childhood programs. Because there is no "periodic table of the elements for social interventions," Lipsey pointed out, classification is a judgment call, and not all analysts will make it in the same way.

A related problem is that, even with any reasonably concise definition of a particular intervention, variability abounds. A statistical test used in meta-analysis, the Q test, is a tool for answering the question of whether or not the between-study variation on the effect sizes for a given outcome is greater than one would expect from the within-study sampling error. Lipsey explained that "it's not unusual to find three, four, five, six, eight, even ten times as much variability across studies [of social interventions] as one would expect just from the sampling error within." This degree of variability—far greater than what is typical in medical studies, for example—is inconsistent with a conclusion that the effects can be generalized. Figure 6-1 illustrates the major sources of variance in studies of social interventions, using the results of an analysis of meta-analyses of psychological, education, and behavioral interventions (Wilson and Lipsey, 2001). The numbers reflect the rough proportions of different sources of variation. Lipsey highlighted how much of the variability is associated with aspects of the methodology—almost as much as is associated with the characteristics of the interventions themselves.

In other words, he noted, "effect size distributions are being driven almost as much by the input of the researchers as they are by the phenomenon that researchers are studying." And this variability may obscure

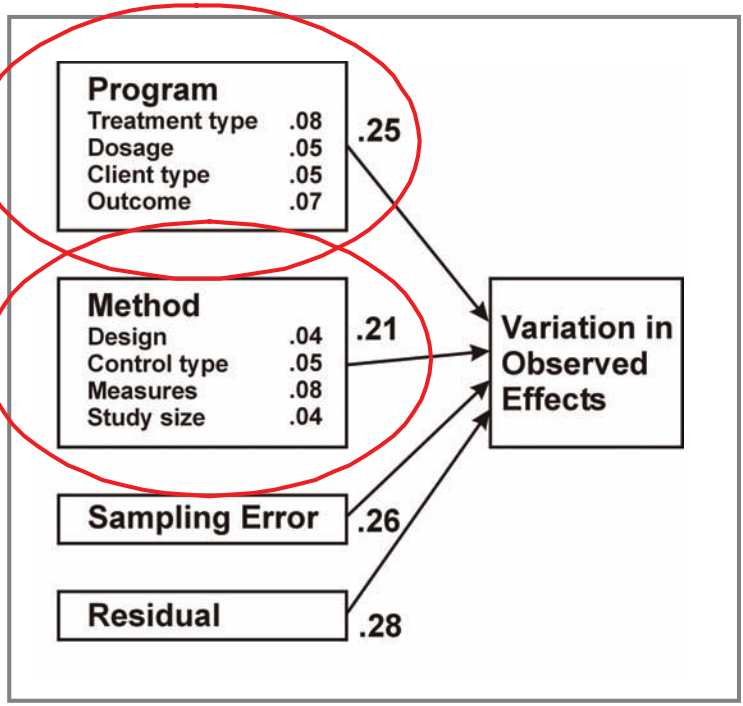


FIGURE 6-1 Many sources of variance in observed effects of social interventions.
NOTE: Estimates based on 300 meta-analyses of intervention studies.
SOURCE: Lipsey (2009).

any real effects. The fact that there is so much variability across studies also highlights important questions about interpreting a single study. The results of single studies are partly a function of the design, procedures, measurement, timing, and so forth; but without the context of other studies, that variability is not evident. Lipsey explained that the methodological variability not only results from differences in design associated with randomization, but also stems from variation in the way outcomes are operationalized. Because there is no settled way of measuring, for example, the noncognitive outcomes of pre-K programs, researchers have been creative, using parent reports, teacher reports, observations, or various standardized scales. They presume that these measures are all targeting the same underlying construct, but there is little evidence to support that presumption.

The biggest problem, however, may be that methodological variation tends to be confounded with substantive variables. Suppose two stud-

ies with different outcomes and samples also show different results that suggest the intervention may work better for African American children than for white children. In fact, those different results may be the result of different measurement procedures or other methodological differences between the two studies and not differential effectiveness of the intervention at all. Since there is significant variation even among studies of a very specific intervention, it is important to consider other factors that may be just as influential on the results.

Still another complication arises because so many studies are small-scale research and development projects, in which the researcher has designed a program and then evaluates it. In practice, there are very few studies in which an independent evaluator studies a real-world program implemented in routine practice. Lipsey reported that his meta-analytical work indicates that research and development studies showed approximately twice the effect size as studies of routine practice. This suggests that one cannot necessarily expect the effects from research and development studies to carry forward when the program is scaled up in a real-world setting.

One implication of this large degree of variation is that the average effect size is unrepresentative of the intervention effects when there is a broad distribution—so variability around the mean effect size is a more useful result to examine. For example, the average effect might be small, but the high end of the distribution could show quite large effect sizes while the low end showed a negative effect. Estimates of average effect sizes are typically used in benefit-cost analysis, so Lipsey urged that they be interpreted with caution.

What can be done about these difficulties? One approach is response surface or, more specifically, effect size surface modeling, an approach for statistically modeling the relationships between intervention effects and key explanatory variables.[1] The response surface is defined by the multiple dimensions of interest along which effects—such as subject characteristics, intervention characteristics, settings, methodology, and the like—vary. Multivariate regression models are used to predict the expected outcomes for a range of scenarios, defined as particular relationships among these varying characteristics. With sufficient studies to map this surface, it is possible to make assumptions about various factors (such as methodology, clientele, setting, and so forth) and estimate the effect size that would be likely in a particular scenario—even if some combinations are not represented in any study.

The result of this sort of analysis is likely to provide a better characterization of intervention effects than a simple mean effect size across avail-

[1]Lipsey credited Donald Rubin for this approach (Rubin, 1990, 1992).

able studies, in Lipsey's view. However, to conduct this analysis requires a relatively large number of diverse studies that provide detailed reporting on the dimensions of interest. Lipsey pointed out that many studies include pages of discussion of methodology, dependent variables, and so forth, with just a few sentences devoted to describing the intervention itself. Nevertheless, this multivariate modeling approach could make it possible to get more out of small, well-executed studies than from larger, more expensive studies that may represent less diversity on key dimensions and be somewhat underpowered, despite their size, for detecting moderator relationships on those dimensions.

DESIGN AND ANALYSIS CONSIDERATIONS

Building on Lipsey's discussion, Howard Bloom addressed a few key issues that arise when researchers attempt to generalize from observed variation in intervention effects. He began with what he described as first principles:

- The purpose of generalizing is to create knowledge by deepening understanding and to inform decisions by projecting findings beyond their immediate context.
- Generalizations project findings to larger populations, sets of outcomes, types of interventions, or types of environments.
- Generalization is done using both statistical sampling and through explanation.

He turned first to the question of how to plan for, analyze, and interpret subgroup findings. The dilemma, he explained, is reconciling different sorts of information into accurate findings that make sense to different audiences. Practitioners treat individuals, policy makers target defined groups, and researchers study averages and patterns of variation—yet all need to learn from the same sources of information. That dilemma translates into questions about how to test multiple hypotheses and identify statistically significant findings. Statistics, he noted, is very limited in its ability to deal with subgroup analysis without using extremely conservative adjustments in advance to avoid Type 1 errors, which tend to wipe out any hints of effects.

One participant noted that results may be very different, depending on whether the analysis uses an interaction model, in which there is a main effect and an interaction effect (as psychologists often prefer) or uses a split sample, and asked whether it would be best to report both. Bloom suggested that the difference is more than technical, that the two methods

actually answer different questions: "Is there a difference among the subgroups?" versus "Is there an interaction effect among several factors?"

Bloom suggested two points to guide effective generalizations. First is specifying subgroups of interest in advance, on the basis of theory, empirical evidence, and policy relevance. Findings that are relevant to subgroups that are defined after the study, based on the results, are exploratory, he suggested, and should be treated differently from findings that confirm (or disconfirm) a hypothesis that was tested. Second, it is important to consider both statistical and substantive significance. There are many ways to assess the statistical significance of findings, looking at the presence of an effect for different subgroups, the size of the effect, and so on. But translating complex findings into substantive messages for nonstatisticians is tricky. For example, small differences between results for subgroups could easily be misinterpreted as suggesting that an intervention was effective for one group and not for the other, when in fact the difference between the groups was not statistically significant.

As an example of a study that successfully modeled variation and effects across subgroups, sites, and studies, Bloom cited one conducted by MDRC of the effects of mandatory welfare-to-work programs for female single parents (Bloom, Hill, and Riccio, 2003). The researchers pulled together data from three MDRC studies to examine the effects of program implementation, the nature of the services offered, client characteristics, and economic conditions. They used a two-level hierarchical model of cross-site variation in experimental estimates of program effects; data covered 59 program offices in 8 states and more than 69,000 participants and included administrative records, participant surveys, and office staff surveys.

The programs studied provided basic education, assistance with job searches, and vocational training. The programs varied in the extent to which they emphasized personal attention to each client and the goal of helping clients secure employment quickly and in other aspects of implementation, and the studies used common measures of these sources of variation. The researchers wanted to find out which characteristics had the biggest impact on short-term outcomes—the outcome measure they used was client earnings during the first two years after they were randomly assigned to receive or not receive the intervention (they also had regional unemployment data for the period studied).

The study had a twofold purpose, however. It was designed not only to build understanding of the relationship between the ways the programs were implemented and their impact, but also to demonstrate a model for generalizing from a range of information. The key findings were the following:

- A strong employment message markedly increased program effects (this was the strongest effect).
- Emphasis on personal client attention increased program effects.
- Large caseloads reduced program effects.
- Reliance on basic education reduced program effects in the short run.
- High unemployment reduced program effects.
- Program effects did not vary consistently with client characteristics.

Perhaps more important, however, is success with a research model that makes use of preplanned subgroup analysis as well as common measures and protocols across studies. Others agreed, suggesting that if some modest core measures for critical outcomes and variables could be established for common use, it would greatly facilitate the work of meta-analysis.

7

Benefit-Cost Analysis in a Policy Context

These methodological and conceptual questions can have a profound influence on policies that affect children and families every day. Rigorous benefit-cost analysis is relatively new in the early childhood context, but available analyses generally point to benefits that significantly outweigh costs. Still, the message to policy makers is not crisp; differences among programs, settings, populations served, goals, available data, and measurement approaches all affect outcomes, costs, and overall conclusions about the value of early childhood programs.

The field faces a double challenge: improving research methods while providing policy makers with accurate information to guide social policy and public investments for children and families. In the final session of the workshop, several views of the tension between research and policy were presented, followed by discussion about future goals and directions.

PERSPECTIVES

Rudy Penner, Jon Baron, and Steve Aos provided three perspectives on the relationship between research and decision making for policy makers.

Keep It Simple

Penner, who offered the perspective of a political veteran, began the discussion with a look at the challenge of using program evalua-

tions as the basis for budget allocations. The fundamental problem, he suggested, is that "all that budgets do is measure the cost of inputs to various programs—they tell you very little about outputs." He pointed out that this is an old problem, citing President Lyndon Johnson's application of analysis that had been used in the Pentagon to evaluate social programs, President Richard Nixon's management by objectives program, and President Jimmy Carter's zero-based budgeting as examples of efforts to bridge the gap. He suggested that none of these efforts was long-lasting or successful because they became overly bureaucratized. It is difficult to make descriptions of complex social programs that deal with human problems fit into neat categories that work across many sectors.

Penner suggested that benefit-cost analysis is difficult even in the context of flood control projects or highway construction, because calculating discount rates for future benefits and costs is never simple, nor is valuing a human life. But evaluating interventions for children is still more complex, and Penner suggested an alternative approach. Instead of providing an empirically supported value for the output of this kind of social program, it might be more useful to simply identify the outcomes and give politicians the responsibility for calculating the program's value. In his view, not only is it the case that many important outcomes cannot be quantified, but also that good and bad outcomes are often comingled. For example, he recalled a Canadian program designed to help unemployed mothers find jobs. Although it was very successful, a collateral result was an increase in problems with their adolescent children, who lost supervision while their mothers were working. For him, identifying the best response to that situation is a social question. He closed by noting that "if you want to influence policy, you really have to try and identify those things that are important to politicians and help them make the kind of value-based tradeoffs that they have to make."

Focus on Finding Effects

Jon Baron focused on evidence of impact as well, but from a somewhat different perspective. A meaningful benefit-cost analysis, he noted, begins with valid evidence of program effects; the next question is whether the benefits of that effect exceed the costs. He suggested that in many fields—in medicine as well as social policy areas—valid evidence of effectiveness is not common. Many widely accepted conclusions about effective programs are based on observational evidence or small randomized trials with short-term follow-up. These programs often show weak effects or no effects when they are evaluated more rigorously.

He described as an exception an example of a nurse home visitation program for poor, mostly single women in their first pregnancy, which

had been subjected to several high-quality evaluations (Olds et al., 1998, 2004, 2007; Luckey et al., 2008). The program provides regular visits during the pregnancy and for the first two years of the child's life. It has been evaluated in three well-implemented randomized trials, which examined different populations and included long-term follow-up. The program demonstrated sizable effects, including—in the study with the longest follow-up—40 to 70 percent reductions in child abuse and neglect and criminal arrests of the children and their mothers by the time the children reached age 15. Based largely on these results, evidence-based home visitation programs are being scaled up; the U.S. Department of Health and Human Services will spend $13.5 million on such home visitation programs in 2009 and the president's fiscal year 2010 budget proposes $8.6 billion over the next decade.

This is the way it's supposed to work, Baron suggested, but there are few such examples. He cited analysis conducted by the Coalition for Evidence-Based Policy suggesting that only 10 to 15 programs across all social policy areas show sizeable, sustained effects in multiple high-quality evaluation studies (he emphasized that a great number of programs show evidence of effectiveness, but in very few does the evidence meet the highest criteria for rigor). Looking at medical examples, he cited a number of seemingly well-supported interventions or findings that later were found to be ineffective or even harmful in well-conducted randomized controlled trials, including

- intensive efforts to lower blood sugar in diabetic patients (increases risk of death),
- hormone replacement therapy for postmenopausal women (increases risk of stroke and heart disease for many women),
- dietary fiber to prevent colon cancer (shown ineffective),
- use of stents to open clogged arteries (shown no better than drugs for most patients),
- having babies sleep on their stomachs (increases risk of sudden infant death syndrome),
- beta-carotene and vitamin E supplements (antioxidants) to prevent cancer (ineffective or harmful),
- oxygen-rich environment for premature infants (increases risk of blindness),
- recent promising AIDS vaccines (found to double risk of AIDS infection), and
- bone marrow transplants for women with advanced breast cancer (ineffective).

He presented a similar list for social policy, of programs that were believed to be effective but later found to have weak or no effect or even adverse effects. These include education programs, such as Upward Bound, federal dropout prevention programs, and a widely used teacher induction program; programs for troubled youth, such as Scared Straight and DARE; and others.

For Baron, the bottom line is that there is a pressing need for research that can accurately identify interventions that work—that have sizable sustained effects and in which "your grandmother would notice the difference in outcomes between the treatment group and the control group." Benefit-cost analyses are valuable for making the case to policy makers for scaling up those programs with sound evidence of effectiveness and for untangling questions about programs that are very costly. But these analyses are best saved for programs that have already been demonstrated to be effective through rigorous evaluations in typical community settings.

Make the Research Work for Policy Makers

Steve Aos illustrated how Washington State produces and uses evidence in policy decision making. The state legislature formed the nonpartisan Washington State Institute for Public Policy to provide analysis of policy options for lawmakers. The institute has become a valuable resource for state legislators, Aos observed, for several reasons. First, it is locally based and closely tied to the community and the lawmaking process. The staff has close working relationships with the lawmakers, and they know exactly which ones to approach on a given issue.

Second, they work in many policy areas. In recent years the institute has examined crime; education, including early childhood education; child abuse and neglect; substance abuse; and mental health, for example. Other states have separate commissions to address different issues, but, in Aos's view, the advantage to the Washington State approach is that the institute has been able to build trust over many years. They draw on information from many sources, often conducting their own meta-analyses, and distill the answers to the precise questions that are current in the state. They present their information in consistent ways (they use a *Consumer Reports*–type format) so it is easy for busy legislators to find what they need and understand the basis for the conclusions.

And the institute has remained scrupulously nonpartisan; Aos observed that benefit-cost analysis has consistently been the most useful tool for helping Democrats and Republicans to identify an approach they can agree on, regardless of the problem. Because the institute has been willing to recommend cutting programs when evidence emerges that they do not work, they have built trust in the evidence-based approach.

Participants followed up on this point, noting that at all levels of government, studies are often used as weapons in strategic conflicts, rather than as factual resources, so the institute approach is designed to move policy makers past that temptation. At the same time, program advocates may be apprehensive about evaluations if they perceive them as political tools or as potentially inaccurate threats to the program's existence. If, instead, evaluation is viewed as a management tool that can identify the most effective aspects of a program, such as Head Start, that has wide political support because of its mission, it may be more politically useful. Yet the fact that policy makers may not always appreciate the subtlety of research findings is a perpetual problem. When results are not a clear-cut "yes, it works" or "no, it doesn't," there is ample room for misrepresentation of results and confusion about their policy implications.

The scale of costs may also affect the nature of the discussion. An expensive early childhood intervention might be unaffordable despite voluminous evidence of long-term benefits, while a low-cost jobs program might make sense even if its outcomes are fairly modest. Aos noted that one of Washington's biggest successes in applying evidence to policy hinged on a question of cost. When the state began questioning its incarceration rate and the high cost of building more and more prisons, it examined the costs of alternative methods of fighting crime. By changing the mix of its crime-fighting resources, it could achieve the same results with less expensive alternatives to prison—while allowing policy makers to maintain their anticrime credentials.

These different perspectives on the role that evidence can and does play in policy discussions led into a wide-ranging discussion of the major ideas that surfaced over the two days.

LOOKING FORWARD

Robert Haveman kicked off the concluding discussion with an overview of key points. First, he noted that a wide range of technical questions is associated with each element in a benefit-cost analysis. Some of the main unsettled points include how best to define an intervention, how to stipulate the elements of benefits and costs, which potential methodological approaches can reliably account for all relevant effects, how to accurately measure all of the important benefits and costs, how to empirically link the intervention to specific impacts, and how to identify shadow values for nonmarket-based benefits. These technical challenges will need to be resolved in order to generalize from available studies to effective policy making.

For Haveman, the big questions to confront in the early childhood intervention area are the following:

1. With a large increase in spending on early childhood interventions on the way, can benefit-cost analysis be used to guide allocations to the most effective programs or types of intervention?
2. Given the current methodological gaps, would it make sense to use expert panels to settle the sorts of technical questions the workshop has highlighted, including questions about measurement and valuation estimates?
3. Is benefit-cost analysis strong enough to guide future policy choices, or should the research and policy community be asking a different question? For example, would it be wiser to focus on monitoring short-run performance to guide the next stage of policy?

Others posed different questions that highlighted the magnitude of the technical challenges. "Are we in a world where scientists can say the money should be spent in the following way to get the biggest bang for our buck, or are we in a world where we should be talking about planned variation and then program evaluation and monitoring?" one asked. For many, the response was clearly the latter. Investments in preschool programs have wide support, for example, but no single particular model has yet been shown to be most effective. As a result, the policy and research focus is turning toward how to structure planned variations that could reveal specific components that would be desirable in a generic model.

Despite the technical challenges, many in the group felt optimistic about the potential for benefit-cost analysis to provide meaningful guidance to future evaluation efforts. At this point, multiple analyses provide valuable information about outcomes as well as costs, even if methodologists still have issues to unravel. The discussion closed with a few thoughts about what would be most useful. First, many thought that a move toward greater standardization in reporting, not only for benefit-cost analysis but also for evaluation in general, would be very useful. "We are not actually in a position to compare the benefits and costs of various programs at this point," one participant suggested, because they have been measured in such different ways and important outcomes have not been captured. Standardization would make it much easier to capture shadow prices and solve other methodological problems. The Washington State model, in which they use the same shadow prices and comparable methods, demonstrates the value of this approach in the policy context. A core set of measures, with common measurement approaches, would improve comparability.

For this sort of research to have real value to policy makers, as one person put it, "the witch doctors have to agree." Policy makers do not care about regression discontinuity or other technical matters, they want accurate, comparable information. Researchers also need to recognize

that events can dictate a need for policy decisions independent of the pace of research. "There's something a little unsatisfying about waiting 40 years and then looking back and saying, what a great program we had in 1963 in Ypsilanti, Michigan, and now here we are wondering what to do today." Policy makers need the information to develop positions they feel comfortable defending. Participants indicated that opportunities exist now to enhance the methodology and to improve certainty about what works, without waiting for the results of expensive longitudinal studies. Improved access to and use of administrative data, including use of these data to project long-term outcomes, for example, was one opportunity cited by several participants in need of further exploration.

However, another participant noted, methodology "may not be the only place where we should be investing time and talent." Although no one at the workshop proposed that a particular methodological approach solves all problems and should be viewed as the state of the art, some benefit-cost analyses do demonstrate benefits that far exceed the costs. "We should also be thinking about where we can't get proof but we can put together good evidence that is not only persuasive to policy makers but will lead us to good policies and good allocation of resources."

FINAL OBSERVATIONS

Federal and state policy makers are showing increased interest in expanding public investments in early childhood interventions. Multiple studies have provided evidence that many such interventions provide long-term benefits for children, their families, and society, but significant questions remain about the extent to which such benefits translate into savings that outweigh the costs of large-scale programs. Improving the quality of evidence that can be used to identify relevant benefits and costs from early childhood interventions will be a valuable asset to policy discussion and support effective policy decisions.

The workshop participants identified multiple technical challenges that deserve attention. While these challenges are daunting, emerging approaches have the potential to significantly enhance the value of these types of analyses in the policy process. The persistent dilemma is how to make immediate decisions about public investments and program priorities with the information at hand while also striving to obtain knowledge through research and evaluation of different program models and policy strategies. Convincing analysis of benefits and costs would provide a guide to the best ways to spend scarce resources for early childhood programs. Methods for conducting the benefit-cost analysis that can provide this kind of evidence are complex in the context of early childhood. However, in a time of limited resources, new collaborative strategies are

emerging that allow researchers, program staff, and policy makers to standardize definitions and measures, to assign explicit values to outcomes and inputs, and to develop other productive approaches for improving benefit-cost methodologies of early childhood interventions.

References

Anderson, M.L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association, 103*(484), 1481-1495.

Aos, S., Miller, M., and Drake, E. (2006). *Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates.* Olympia: Washington State Institute for Public Policy.

Barnett, W.S. (2009). *Sins of the fathers? Lessons from the Perry, Abecedarian, and CPC studies*. Presentation at the Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, March 4-5, National Academies, Washington, DC. Available: http://www.bocyf.org/barnett_presentation.pdf (accessed 9/15/2009).

Barnett, W.S., and Masse, L.N. (2007). Early childhood program design and economic returns: Comparative benefit-cost analysis of the Abecedarian program and its policy implications. *Economics of Education Review, 26*(1), 113-125.

Belfield, C. (2009). *Costs analysis for early childhood interventions: Evidence from New Jersey*. Presentation at the Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, March 4-5, National Academies, Washington, DC. Available: http://www.bocyf.org/belfield_presentation.pdf (accessed 9/15/2009).

Belfield, C., Nores, M., Barnett, W.S., and Schweinhart, L.J. (2006). The High/Scope Perry Preschool Program: Cost-benefit analysis using data from the age-40 followup. *Journal of Human Resources, 41*(1), 162-190.

Black, S.E., Devereux, P.J., and Salvanes, K.G. (2007). From the cradle to the labor market? The effect of birth weight on adult outcomes. *Quarterly Journal of Economics*, *122*(1), 409-439.

Bloom, H.S., Hill, C.J., and Riccio, J. (2003) Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management, 22*(4), 551-575.

Brooks-Gunn, J., McCarton, C.M., Casey, P.H., McCormick, M.C., Bauer, C.R., Bernbaum, J.C., Tyson, J., Swanson, M., Bennett, F.C., Scott, D.T., Tonascia, J., and Meinert, C.L. (1994). Early intervention in low-birth-weight premature infants: Results through age 5 years from the infant health and development program. *Journal of the American Medical Association, 272*(16), 1257-1262.

Brooks-Gunn, J., Magnuson, K., and Waldfogel, J. (2009). *Long-run economic effects of early childhood programs on adult earnings.* Issue Paper #12. Washington, DC: Partnership for America's Economic Success.

Chazan-Cohen, R., Ayoub, C., Pan, B.A., Roggman, L., Raikes, H., McKelvey, L., Whiteside-Mansell, L., and Hart, A. (2007). It takes time: Impacts of Early Head Start that lead to reductions in maternal depression two years later. *Infant Mental Health Journal, 29*(2), 151-170.

Chicago Longitudinal Study. (2004). *Chicago longitudinal study*. Available: http://www.waisman.wisc.edu/cls/ (accessed 9/15/2009).

Chicago Public Schools. (2009). *Child Parent Center (CPC)*. Available: http://www.cps.edu/Schools/Preschools/Pages/Childparentcenter.aspx (accessed 9/15/2009).

Cohen, M.A., Rust, R.T., Steen, S., and Tidd, S. (2004). Willingness-to-pay for crime control programs. *Criminology, 42*(1), 86-106.

Cook, P.J., and Ludwig, J. (2000). *Gun violence: The real costs.* New York: Oxford University Press.

Currie, J., and Thomas, D. (1995). Does Head Start make a difference? *American Economic Review, 85*(3), 341-364.

Currie, J., Stabile, M., Manivong, P., and Roos, L.L. (2008). *Child health and young adult outcomes.* NBER Working Paper No. 14482. Cambridge, MA: National Bureau of Economic Research.

Dehejia, R.H., and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151-161.

Doyle, J.J., Jr. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy, 116*(4), 746-770.

Education Law Center. (2007). *Abbott Preschool Program*. Available: http://www.edlawcenter.org/ELCPublic/AbbottPreschool/AbbottPreschoolProgram.htm (accessed 9/15/2009).

FPG Child Development Institute. (2009a). *The Carolina Abecedarian Project: Major findings.* Chapel Hill: University of North Carolina. Available: http://www.fpg.unc.edu/~abc/#major_findings (accessed 9/15/2009).

FPG Child Development Institute. (2009b). *Environment rating scales.* Chapel Hill: University of North Carolina. Available: http://www.fpg.unc.edu/~ecers/ (accessed 9/15/2009).

Garces, E., Thomas, D., and Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review, 92*(4), 999-1012.

Glotzer, D.E., Freedburg, K., and Bauchner, H. (1995). Management of childhood lead poisoning: Clinical impact and cost-effectiveness. *Medical Decision-Making*, *15*(1), 13-23.

Gormley, W.T., Gayer, T., Phillips, D., and Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872-884.

Gray, S.W., and Klaus, R.A. (1970). The Early Training Project: A seventh-year report. *Child Development*, *41*(4), 909-924.

Haveman, R.H., and Wolfe, B.L. (1984). Schooling and economic well-being: The role of nonmarket effects. *Journal of Human Resources, 19*(3), 377-407.

HighScope Educational Research Foundation. (2009). *HighScope Perry Preschool Study*. Available: http://www.highscope.org/Content.asp?ContentId=219 (accessed 9/15/2009).

Hopkins, R.B., Paradis, J., Roshankar, T., Bowen, J., Tarride, J.-E., Blackhouse, G., Lim, M., O'Reilly, D., Goeree, R., and Longo, C.J. (2008). Universal or targeted screening for fetal alcohol exposure: A cost-effectiveness analysis. *Journal of Studies on Alcohol and Drugs, 69*(4), 510-519.

Institute of Medicine. (2006). *Valuing health for regulatory cost-effectiveness analysis*. Committee to Evaluate Measures of Health Benefits for Environmental, Health, and Safety Regulation, W. Miller, L.A. Robinson, and R.S. Lawrence (Eds.). Board on Health Care Services. Washington, DC: The National Academies Press.

Karoly, L.A. (2009). *Overview of benefit-cost analysis for early childhood interventions*. Presentation at the Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, March 4-5, National Academies, Washington, DC. Available: http://www.bocyf.org/karoly_presentation.pdf (accessed 9/15/2009).

Karoly, L.A., Kilburn, M.R., and Cannon, J.S. (2005). *Early childhood interventions: Proven results, future promise*. Arlington, VA: RAND Corporation.

Kennedy, M.M. (1978). Findings from the Follow Through Planned Variation Study. *Educational Researcher, 7*(6), 3-11.

Kilburn, M.R., and Karoly, L.A. (2008). *The economics of early childhood policy: What the dismal science has to say about investing in children.* Labor and Population Occasional Paper. Arlington, VA: RAND Corporation.

Krueger, A.B. (2003). Economic considerations and class size. *Economic Journal*, *113*(485), F34-F63.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*(4), 604-620.

Lipsey, M.W. (2009). *Generalizability: The role of meta-analysis*. Presentation for the Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, March 4-5, National Academies, Washington, DC. Available: http://www.bocyf.org/lipsey_presentation.pdf (accessed 9/15/2009).

Luckey, D.W., Olds, D.L., Zhang, W., Henderson, C., Knudtson, M., Eckenrode, J., Kitzman, H., Cole, R., and Pettitt, L. (2008). *Revised analysis of 15-year outcomes in the Elmira Trial of the Nurse-Family Partnership*. Prevention Research Center for Family and Child Health, University of Colorado Department of Pediatrics.

Ludwig, J., and Miller, D.L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity approach. *Quarterly Journal of Economics, 122*(1), 159-208.

Magnuson, K. (2009). *Approaches to projecting (or guesstimating) long-term outcomes*. Presentation for the Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions, March 4-5, National Academies, Washington, DC. Available: http://www.bocyf.org/magnuson_presentation.pdf (accessed 9/15/2009).

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., and Leibowitz, A. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review, 77*(3), 251-277.

Masse, L.N., and Barnett, W.S. (2002). *A benefit cost analysis of the Abecedarian early childhood intervention*. New Brunswick, NJ: National Institute for Early Education Research.

McCarton, C.M., Brooks-Gunn, J., Wallace, I.F., Bauer, C.R., Bennett, F.C., Bernbaum, J.C., Broyles, S., Casey, P.H., McCormick, M.C., Scott, D.T., Tyson, J., Tonascia, J., and Meinert, C.L. (1997). Results at age 8 years of early intervention for low-birthweight premature infants. *Journal of the American Medical Association, 277*(2), 126-132.

McCormick, M.C., Brooks-Gunn, J., Buka, S.L., Goldman, J., Yu, J., Salganik, M., Scott, D.T., Bennett, F.C., Kay, L.L., Bernbaum, J.C., Bauer, C.R., Martin, C., Woods, E.R., Martin, A., and Casey, P. (2006). Early intervention in low birthweight premature infants: Results at 18 years of age for the infant health and development program. *Journal of Pediatrics, 117*(3), 771-780.

Mishan, E.J. (1971). Evaluation of life and limb: A theoretical approach. *Journal of Political Economy, 79*(4), 687-705.

National Head Start Association. (2009). *National Head Start Association*. Available: http://www.nhsa.org/ (accessed 10/20/2009).

National Institute for Early Education Research. (2006) *The state of preschool: 2005 state preschool yearbook.* Washington, DC: Pew Center on the States. Available: http://www.pewcenteronthestates.org/report_detail.aspx?id=32904 (accessed 10/20/2009).

National Research Council and Institute of Medicine. (2000). *From neurons to neighborhoods: The science of early childhood development.* Committee on Integrating the Science of Early Childhood Development, J.P. Shonkoff and D.A. Phillips (Eds.). Board on Children, Youth, and Families, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council and Institute of Medicine. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities.* Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth, and Young Adults: Research Advances and Promising Practices, M.E. O'Connell, T. Boat, and K.E. Warner (Eds.). Board on Children, Youth, and Families, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olds, D.L., Henderson, C.R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., and Powers, J. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Journal of the American Medical Association, 280*(14), 1238-1244.

Olds, D.L., Kitzman, H., Hanks, C., Cole, R., Anson, E., Sidora-Arcoleo, K., Luckey, D.W., Henderson, C.R., Holmberg, J., Tutt, R.A., Stevenson, A.J., and Bondy, J. (2007). Effects of nurse home visiting on maternal and child functioning: Age-9 follow-up of a randomized trial. *Pediatrics, 120*, e832-e845.

Olds, D.L., Robinson, J., Pettitt, L., Luckey, D.W., Holmberg, J., Ng, R.K., Isacks, K., Sheff, K., and Henderson, C.R. (2004). Effects of home visits by paraprofessionals and by nurses: Age 4 follow-up results of a randomized trial. *Pediatrics, 114*(6), 1560-1568.

Puma, M., Bell, S., Cook, R, Heid, C., Lopez, M., et al. (2005). *Head Start impact study: First year findings.* Westat. Report prepared for the U.S. Department of Health and Human Services.

Rubin, D.B. (1990). A new perspective on meta-analysis. In K.W. Wachter and M.L. Straf (Eds.), *The future of meta-analysis* (pp. 155-165). New York: Russell Sage Foundation.

Rubin, D.B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics, 17*(4), 363-374.

Schwartz, J., Pitcher, H., Levin, R., Ostro, B., and Nichols, A.L. (1985). *Costs and benefits of reducing lead in gasoline: Final regulatory impact analysis.* Washington, DC: U.S. Environmental Protection Agency.

Smith, J.P. (2007). The impact of socioeconomic status on health over the life-course. *Journal of Human Resources*, *42*(4), 739-764.

Taylor, L.L., and Fowler, W.J., Jr. (2006). *A comparable wage approach to geographic cost adjustment* (NCES-2006-321). National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Temple, J.A., and Reynolds, A.J. (2007). Benefits and costs of investments in preschool education: Evidence from the Child-Parent Centers and related programs. *Economics of Education Review, 26*(1), 126-144.

Tolley, G.S., Kenkel, D.S., and Fabian, R.G. (1994). *Valuing health for policy: An economic approach*. Chicago: University of Chicago Press.

U.S. Environmental Protection Agency. (1997). *The benefits and costs of the Clean Air Act, 1970 to 1990.* Washington, DC: Author.

Weimer, D.L., and Vining, A.R. (2009). *Investing in the disadvantaged: Assessing the benefits and costs of social policies*. Washington, DC: Georgetown University Press.

Wilson, D.B., and Lipsey, M.W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413-429.

Wolfe, B.L., and Haveman, R. (2001). Accounting for the social and nonmarket benefits of education. In J.F. Helliwell (Ed.), *The contribution of human and social capital to sustained economic growth and well-being: International symposium report*. OECD/Human Resources Development Canada. Vancouver: University of British Columbia Press.

Zhai, F., Brooks-Gunn, J., and Waldfogel, J. (2009). *The effects of Head Start on the school readiness of children in fragile families.* Fragile Families Working Group Seminar, Columbia Population Research Center, February 12. Available: http://cupop.columbia.edu/node/311 (accessed 9/15/2009).

Appendix A

Glossary

attrition—in the context of research studies, refers to the gradual loss of study participants, some percentage of whom often drop out.

benefit-cost analysis—a method of economic analysis in which both costs and outcomes of an intervention are valued in monetary terms, permitting a direct comparison of the benefits produced by the intervention with its costs (also referred to as cost-benefit analysis).

contingent valuation analysis—a method of obtaining estimates of the worth of a social good or benefit in which people are asked how much they would pay for a particular outcome, given a particular hypothetical scenario.

cost-effectiveness analysis—a method of economic analysis in which outcomes of an intervention are measured in nonmonetary terms. The outcomes and costs are compared with both the outcomes (using the same outcome measures) and the costs for competing interventions, or with an established standard, to determine if the outcomes are achieved at reasonable monetary cost.

dependent variable—the factor(s) that change as a result of an experimental treatment or intervention, such as, for example, the academic skills of children who have participated in an early childhood education program.

discount rate—a factor used to estimate future costs or the value of future benefits at the current equivalent value, used with the goal of attempting to take into account likely changes in valuation, opportunity costs, and other factors.

economy of scale—advantages that accrue when a project is conducted on a larger scale than initially, which result from opportunities to use resources more efficiently and to reduce costs.

effect size—the magnitude of results (or effects on participants) of a particular treatment or intervention that is being studied.

independent variable—one of the characteristics of an experiment's subjects that are considered in the study design, such as, for example, the age and gender of the participants in an early childhood program.

intent-to-treat—the group of study participants randomly selected to receive the intervention being studied.

multivariate regression model—a statistical procedure for examining experimentally the relationship among several variables. By making it possible to distinguish the impact on outcomes of one variable from the impacts of others, this analysis makes it possible to control for factors that may influence the results and obscure the effects the experiment is intended to identify.

opportunity cost—the value of alternatives not chosen, calculated as part of an analysis of the costs of the alternative that was chosen.

plug-in—estimates for particular costs that can be used to streamline cost analysis.

p-value—calculation of the probability that the data indicate a significant difference.

quasi-experimental design—an experiment designed to produce evidence of causality when randomized controlled trials are not possible, using alternative statistical procedures to compensate for nonrandom factors.

randomized controlled trial—an experiment in which the participants are assigned by chance either to receive the intervention or treatment being studied or not to receive it, so that the results can be compared across

statistical identical groups. When this is done with a large enough number of participants, any differences among them that might influence their response to the treatment will be distributed evenly.

regression adjustment—a statistical technique for reducing bias in an experiment that can occur when variables other then the one(s) being studied may affect the results in nonrandom ways.

regression discontinuity design—a quasi-experimental analysis that can be used in program evaluation when randomized assignment is not feasible. It is based on the assumption that individuals who fall just above or below a cut-off point on a particular scale are likely to be similar, so that this group can be treated as varying randomly.

selection bias—an unrecognized difference in the characteristics of the subjects of an experiment who do or do not receive the treatment, or who or do not benefit from it, that will affect the results.

shadow value/shadow price—the true value or cost of the results of a particular decision, as calculated when no market price is available; a dollar value attached to an opportunity cost.

worst-case bounds—a statistical analysis in which the outer limit assumptions for an experiment—both the best possible and worst possible outcomes in terms of the data supporting or not supporting the experimental hypothesis-—are examined. This analysis provides a way of assessing the significance of actual error that may occur in any experiment.

Appendix B

Workshop Agenda and Participants

AGENDA

Workshop on Strengthening Benefit-Cost Methodology for the Evaluation of Early Childhood Interventions

Wednesday, March 4, 2009

12:30-12:45 PM Welcoming Remarks
Barbara L. Wolfe, University of Wisconsin-Madison and Planning Committee Chair

12:45-1:15 PM Introduction: Overview of Benefit-Cost Analysis for Early Childhood Interventions
Lynn A. Karoly, RAND Corporation

1:15-2:30 PM *Panel 1*: Methodological Issues in Evaluation Design
Moderator: *Jane Waldfogel, Columbia University*

Identification
Jens Ludwig, University of Chicago

Statistical Inference
David Deming, Harvard University

Questions and Discussion

2:30-3:30 PM *Panel 2*: Resources and Costs for Full-Scale Early Childhood Interventions
Moderator: *David L. Weimer, University of Wisconsin-Madison*

Issues in Assessing Costs and Determining Resource Allocations
Henry Levin, Columbia University

An Illustration of Cost Estimation from New Jersey
Clive Belfield, Queens College, CUNY

Questions and Discussion

3:30-4:45 PM *Panel 3*: Early Childhood Intervention Outcomes for Benefit-Cost Analysis
Moderator: *Margaret C. Simms, The Urban Institute*

Sins of the Fathers
W. Steven Barnett, Rutgers, The State University of New Jersey

The Wish List
Jeanne Brooks-Gunn, Columbia University

Questions and Discussion

4:45-5:00 PM Closing Remarks and Adjournment
Barbara L. Wolfe

Thursday, March 5, 2009

8:30-8:40 AM Welcoming Remarks
Barbara L. Wolfe

8:40-10:00 AM *Panel 4*: Assessing Long-Term Outcomes
Moderator: *Jane Waldfogel*

Approaches to Projecting Long-Term Outcomes
Katherine A. Magnuson, University of Wisconsin-Madison

Leveraging Administrative Data
Janet Currie, Columbia University

Questions and Discussion

10:00 AM-12:00 PM *Panel 5*: Valuation of Outcomes and Resources/Costs
Moderator: *Robert M. Kaplan, University of California, Los Angeles*

Shadow Prices Needed for CBAs of Early Childhood Interventions
David L. Weimer

Valuation of Outcomes in Environmental Economics
Myrick Freeman, Bowdoin College

Valuation of Outcomes in Criminal Justice
Philip J. Cook, Duke University

Valuation of Outcomes in Health Economics
Donald S. Kenkel, Cornell University

Questions and Discussion

12:00-12:45 PM Lunch

12:45-2:10 PM *Panel 6*: Generalizability of Benefit-Cost Analyses
Moderator: *Barbara L. Wolfe*

The Potential Role of Meta-Analysis
Mark W. Lipsey, Vanderbilt University

Design and Analysis Considerations
Howard S. Bloom, MDRC

Questions and Discussion

2:10-3:30 PM	*Panel 7*: Policy Decision-Making Roundtable Moderator: *Ron Haskins, The Brookings Institution* Roundtable Panel Members *Rudolph G. Penner, The Urban Institute* *Jon Baron, Coalition for Evidence-Based Policy* *Steve Aos, Washington State Institute for Public Policy*
3:30-4:30 PM	Workshop Wrap-Up Moderator: *Robert M. Kaplan* Summary Comments: *Robert H. Haveman, University of Wisconsin-Madison* Discussion
4:30-5:00 PM	Final Participant Comments and Closing Remarks *Barbara L. Wolfe*

PARTICIPANTS

Committee Members:

Barbara L. Wolfe (*Chair*), Department of Population Health Sciences and Department of Economics, University of Wisconsin-Madison
Ron Haskins, Economic Studies Program, The Brookings Institution
Robert M. Kaplan, School of Public Health, University of California, Los Angeles
Lynn A. Karoly, RAND Corporation
Henry M. Levin, Teachers College, Columbia University
Jens Ludwig, Harris School of Public Policy Studies, University of Chicago
Margaret C. Simms, The Urban Institute
Jane Waldfogel, Radcliffe Institute for Advanced Studies, Harvard University and Columbia University
David L. Weimer, Robert M. Lafollette School of Public Affairs, University of Wisconsin-Madison

Workshop Presenters:

Steve Aos, Washington State Institute for Public Policy
W. Steven Barnett, National Institute for Early Education Research, Rutgers, The State University of New Jersey
Jon Baron, Coalition for Evidence-Based Policy
Clive Belfield, Economics Department, Queens College, CUNY
Howard S. Bloom, MDRC
Jeanne Brooks-Gunn, Teachers College, Columbia University
Philip J. Cook, Sanford Institute of Public Policy, Duke University
Janet Currie, Economics Department, Columbia University
David Deming, Kennedy School of Government, Harvard University
A. Myrick Freeman, Department of Economics, Bowdoin College
Robert H. Haveman, Institute for Research on Poverty, University of Wisconsin-Madison
Donald S. Kenkel, Department of Policy Analysis and Management, Cornell University
Mark W. Lipsey, Institute of Public Policy, Vanderbilt University
Katherine A. Magnuson, School of Social Work, University of Wisconsin-Madison
Rudolph G. Penner, The Urban Institute

National Academies Staff:

Alexandra Beatty, Rapporteur
Barbara Boyd, Administrative Coordinator
Rosemary Chalk, Director, Board on Children, Youth, and Families
Wendy Keenan, Program Associate
Bridget Kelly, Program Officer
David Myles, Christine Mirzayan Science and Technology Policy Fellow
Mary Ellen O'Connell, Study Director
Julienne Palbusa, Research Assistant

Registered Attendees:

Nobel Absalom, Congressional Budget Office
Douglas Almond, Columbia University
Jill Antonito, Pew Charitable Trusts
Katherine Astrich, Office of Management and Budget
Tim Bartok, Upjohn Institute for Employment Research
Stephen Bell, Bat Associates Inc.
Bond Benton, My Way Foundation
Dara Blachman, Federal Interagency Forum on Child and Family Statistics

Melissa Broods, Children's Bureau
Jennifer Brooks, Administration for Children and Families
Ajay Chuddy, The Urban Institute
Dale Church, Dale Church Consulting Inc.
Rachel Cohen, Administration for Children and Families
Laura Dinehart, Florida International University
Kathleen Dwyer, Administration for Children and Families
Daniel Eisenberg, University of Michigan
Curtis Florence, Centers for Disease Control and Prevention
Karen Freely, Ounce of Prevention Fund
Sarah Friedman, CNA
William Gormley, Georgetown University
Daryl Greenfield, University of Miami
James Griffin, National Institute of Child Health and Human Development, National Institutes of Health
Scott Grosse, Centers for Disease Control and Prevention
Robert Grunewald, The Federal Reserve Bank of Minneapolis
Emily Holcombe, Center for Research on Children in the United States, Georgetown University
Allison K. Holmes, Administration for Children and Families
Julia Isaacs, Brookings Institution
Alex Kemper, Duke University
Rebecca Kilburn, RAND Corporation
Julie Lee, Congressional Budget Office
Vicky Marchland, The Finance Project
Nancy Geyelin Margie, Administration for Children and Families
Ivelisse Martinez-Beck, Administration for Children and Families
Karen Matsuoka, Office of Management and Budget
John McCoy, Fight for Children
Song Hayek Moon, University of Chicago
Martha Morehouse, Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services
Robert Palmer, Dale McMurchy Consulting Inc.
Patricia Pastor, Centers for Disease Control and Prevention
Ruth Perou, Centers for Disease Control and Prevention
Robin Pulliam, U.S. House Committee on Education and Labor
David Racine, University of Illinois at Urbana-Champaign
Dan Rosenbaum, Office of Management and Budget
Christine Ross, Mathematica Policy Research
Erin Schelar, Child Trends, Inc.
Lisbeth Schorr, Center for the Study of Social Policy
Keith Scott, University of Miami
Heather See, University of Maryland, College Park

Ann Segal, Wellspring Advisors
Joan Smith, Casey Family Programs
Michael Stoto, Georgetown University
Louisa Tarullo, Mathematica Policy Research
Melissa Thornton, Public Health Agency of Canada
William Turner, University of Minnesota
Jennifer Urban, Office of Behavioral and Social Sciences Research, National Institutes of Health
Kristin Ward, Casey Family Programs
Albert Wat, Pew Charitable Trusts
Sara Watson, Pew Charitable Trusts
Mary Bruce Webb, Administration for Children and Families
Michael Weinstein, Robin Hood Foundation
Elaine Weiss, Pew Center on the States
T'Pring R. Westbrook, Administration for Children and Families
Richard Zerbe, University of Washington