

Understanding Evaluations of Home Visitation Programs

Deanna S. Gomby

Abstract

This journal issue comprises reports concerning program evaluations of key national home visitation models. No single evaluation can answer all the questions of interest about a program, nor is any evaluation perfect, which means that readers must carefully weigh the intended purpose of the evaluation and the evaluation's strengths and weaknesses before deciding what conclusions can credibly be drawn from its results.

This article begins with a discussion of the role of evaluation both in improving programs and in determining program effects. The choices required to craft a strong and methodologically rigorous evaluation are described: what outcomes to measure and how; what methods to use in designing the evaluation and building a comparison group; how many participants to enroll; and how to devise a strong plan for data analysis involving subgroups of the enrolled families.

The article then discusses additional factors policymakers and practitioners should consider when interpreting the results of home visiting evaluations: attrition, the policy and functional importance of the outcomes, and the likely generalizability of the results to other communities or other populations.

The evaluations that appear in this journal issue are used as examples throughout the article, and the measures that were used in those evaluations are summarized. The evaluations included in this journal issue have both strengths and weaknesses but are probably among the better evaluations in the home visiting field.

Deanna S. Gomby, Ph.D., is deputy director of Children, Families, and Communities at The David and Lucile Packard Foundation.

The Purposes of Program Evaluation

Evaluations of human-service programs are typically designed to answer one or more of the following questions:

- (1) What services did the program provide?
- (2) Who received the services?
- (3) Did the services produce the anticipated outcomes?

If the primary purpose of the evaluation

is to help program staff hone a new program, then answering the first two questions may be enough. If, on the other hand, the purpose of the evaluation is to persuade funders to continue or expand support or to sponsor replication in other communities, then the question about program outcomes must be answered.

These questions are not easily separable, of course, and most of the evaluations in this journal issue include information designed to answer all three questions. As the articles

demonstrate, however, answering them can be difficult. This article describes some of the reasons why these questions are important for policymakers and practitioners, why they are hard to answer, and how the evaluations in this journal chose to address them.

What Services Did the Program Provide?

Answering this question can provide information that can be used both to improve a program and to interpret the results of evaluations focused on program effectiveness. For example, if an evaluation suggests that services are not being delivered as intended, then program administrators may want to institute quality-improvement measures to improve implementation, or they may decide that the model or curriculum should

reach families. Alternatively, perhaps the planned intensity level of services is simply unrealistic. Or perhaps the model needs to be modified to make it more interesting, and then parents will seek it out more readily.

No matter whether the results are used to improve existing practice or to alter the model, understanding these variations in “dosage” can have implications for understanding the eventual outcomes of any program. Some of the evaluations (for example, see the article by Wagner and Clayton in this journal issue) suggest that families that receive higher-intensity services benefit more than those that receive fewer visits; if this is correct, then knowing that the tested intervention is not delivering as many visits as planned may mean that the program will be less likely to produce the intended benefits.

For many families, the intervention that was tested was not as intensive an intervention as the model developers planned.

be modified because practice in the field is suggesting a better approach. Program evaluators may use implementation information to make sure that the evaluation is a fair test of the intended intervention and not an evaluation of a poorly implemented shadow. In addition, evaluators can use implementation information to explain the results that their evaluations of outcomes eventually produce.

Intensity of Services

Several of the reports in this journal issue suggest that families received fewer home visits than were intended by their models—in some cases, families averaged about 40% to 60% of the number of visits intended in the models (among those reporting this information were Parents as Teachers [PAT] and the Nurse Home Visitation Program). For many families, therefore, the intervention that was tested was not as intensive an intervention as the model developers planned.

That information by itself cannot determine the next steps, but it can alert program planners to some possible alternatives. For example, it might suggest that visitors need additional training in how to contact hard-to-

Content of the Visits

The content of the home visits may also stray from the intended curriculum. Most of the home visitation programs described in this journal issue have core curricula, but visitors may not always be able to deliver the lesson plans. A mother may be concerned about a sick infant, or may have had a very rough night with an abusing spouse, and she may want to talk about those issues rather than about the presumed topic for the day. The home visitor is likely to set aside the curriculum to address the mother’s more pressing concerns. That ability to respond to parental concerns immediately and with sensitivity is one of the hallmarks of home visitation programs, and is widely seen as one of their strengths. Nevertheless, if such deviation occurs on a regular basis, or if individual home visitors consistently vary their programs as a reflection of their own backgrounds and experiences, then the service the home visitors provide is not the same as what program designers originally proposed.

The evaluations in this journal issue do not directly report on this aspect of program implementation, although Baker, Piotrkowski, and Brooks-Gunn suggest in their article about the Home Instruction Program for Preschool Youngsters (HIPPY) that variation in delivery of the intended curriculum does occur. Usually, evaluators try to capture the content of the services through (1) interviews with home visitors conducted some time after the visits occur (see the arti-

cle by Baker, Piotrkowski, and Brooks-Gunn in this journal issue), (2) reports by home visitors summarizing what occurs during the lessons, or (3) results of analyses of videotapes of actual home visits (see the article by Wagner and Clayton in this journal issue).

If evaluations reveal that differences in program content have occurred, program planners may want to change the model to incorporate the changes the home visitors are making. Or, if they believe the differences reflect poor training, they may institute in-service training or closer supervision to encourage more faithful implementation.

From a methodological point of view, however, averaging results across all the home visitors in a particular program, with their own styles and session content, may disguise the differences present, and so mask program effectiveness. Because such individualization of services is inherent in home visiting, it is quite possible that this has occurred to some extent in all the evaluations reported in this journal issue. Only a careful analysis of information concerning what actually occurred during home visits would allow this to be disentangled, and such information is not available for most programs.

Ancillary Services

The services that are provided in the home are often only a part of the total intervention. Some programs (for example, HIPPI and PAT) offer both home visits and parent group meetings. The HIPPI evaluation reported that some types of families were more likely to attend the group meetings than to persevere with the home visits. If outcomes such as children's development differed among these families, then knowing who actually made use of the offered services might help explain those results.

Most programs also seek to connect families with a range of services in the community, including health and child care services for the children, and employment, housing, transportation, and drug-treatment services for the parents. The services families are referred to and receive essentially become part of the program for those families, and may become a critical element in the observed success or failure of the home visiting program. Evaluators therefore some-

times track the community services families receive (for example, see the article by St.Pierre and Layzer on the Comprehensive Child Development Program [CCDP]), but this is an expensive task that requires the cooperation of participating families and community agencies to either complete interviews or approve the release of family records. For these reasons, few evaluations, including those in this journal issue, capture those ancillary services in great detail.

Determining which services families receive may have implications for program quality: If program administrators believe that the strength of their model depends upon linkages of families with other community institutions, but those linkages never occur, then the administrators may seek other strategies to forge those connections.

The services families are referred to and receive essentially become part of the program and may become a critical element in the observed success or failure of the home visiting program.

From a methodological point of view, the variability in the extent to which families seek out and access services may decrease the likelihood that an evaluation will detect a difference in overall outcomes for all families. In addition, a model that relies heavily on other community services may suggest that the model's success in one community will not necessarily translate to another community.

Who Received the Services?

Describing the people who participated in a program—both those designated as eligible for the program and those who made use of available services—is important for improving the program, for interpreting the results of evaluations, and for making judgments about who should receive services in the future.

Typically, information about program participants begins with information about the eligible population: all mothers with

newborns in the community, or just teens, first-time mothers, or low-income families, and so on. The studies in this journal issue essentially all focused on low-income families, although some programs were offered fairly universally to everyone within a geographic catchment area (for example, PAT), or screened everyone within that area for services, and then offered services to the most needy families (for example, Healthy Families America [HFA] and Hawaii Healthy Start). Although not all of the programs reported information about this issue, those that did (for example, CCDP, Hawaii Healthy Start, and the Nurse Home Visitation Program) suggested that perhaps 10% to 25% of families who are invited to enroll in services refuse to do so.

The next step is to describe who actually received services. The evaluations of HIPPY, Hawaii Healthy Start, and HFA suggest that some types of families are more likely than others either to use some aspects of the programs or to continue participation.

If understood, these differences can suggest program improvements. For example, the article by Baker, Piotrkowski, and Brooks-Gunn in this journal issue suggests that the HIPPY model should be extended so that programs offer services to overcome barriers that prevent families from taking

What is easiest to document in terms of time and cost may not be the most meaningful or the most accurate measure.

advantage of existing services. Policymakers can use this information to judge whether a program, when extended to a new community, is likely to have the same effects. And evaluators can use information about program enrollment and participation to explain changes that the program created (or, perhaps, failed to create) among different groups of participants.

Did the Services Produce the Anticipated Outcomes?

Answering the first two questions—what services were provided and who received them—is a key step in the evaluation of any

service program but is not sufficient to reach conclusions about a home visitation program's effects on children and families. For that, another type of evaluation is required—one that seeks to demonstrate that hoped-for changes in program participants, their families, or their communities have occurred, and that the changes were caused by the program and not by something else.

The extent to which the causal connection can be made is largely determined by several key decisions about the design of the evaluation. These include which outcomes will be assessed and how they will be measured, whether and how a comparison group will be constructed, and how many people will be assessed.

Choosing Outcomes

Ideally, program and evaluation planners should select outcomes carefully based on their implicit and explicit theories of how the program services are supposed to create change, but, in fact, outcomes are usually selected for measurement through a combination of theory and pragmatism.

For example, a home visitation program that is supposed to prevent child abuse and neglect may be hypothesized to work in one or more ways: (1) by increasing parental knowledge or altering parental expectations about child development, (2) by changing parental attitudes toward child rearing, (3) by modifying the parent-child interaction, and/or (4) by increasing surveillance that either leads to earlier detection of potential problems or discourages the expression of those problems. Although a change in rates of abuse and neglect is the ultimate goal for the program, accurately measuring such change is difficult for a variety of reasons, including the general reluctance of parents and others to report abuse and neglect and the wide variability in child protective services agencies' responses to those reports. (See the articles in this journal issue by Duggan and colleagues, by Olds and colleagues, and by Daro and Harding for discussions of this topic.) In other words, assessing changes in child abuse and neglect rates alone may make the program look less effective than it really is and may also mean that evaluators miss changes in intermediate outcomes,

such as parent-child interaction, that are due to the program.

Using a theory to guide the choice of outcomes ensures the assessment of intermediate outcomes all along the hypothesized causal chain and can increase confidence that the program is generating a plausible pattern of results. For example, if no differences in intermediate outcomes are found, then it is less likely that differences will be produced in the outcome at the end of the causal chain—whether or not that outcome is actually measured. If a benefit in the outcome at the end of the chain is present but none of the intermediate benefits was produced, then one might look carefully at the results concerning the ultimate outcome to make sure that they seem plausible.

With such a causal chain approach toward evaluation, programs that focus on child abuse prevention might measure changes in parent knowledge and attitudes, in parent-child relationships, in rates of emergency room visits for injuries and ingestions (which may reflect physical abuse or neglect), and in child abuse and neglect rates (both reports of suspected maltreatment and confirmed incidents).

For most of the home visitation programs discussed in this journal issue, a wide range of outcomes were assessed, some of which are listed in Table 1. These usually focused on outcomes associated with child health and development, including child abuse and neglect; parenting skills, parent behavior, or parent-child interaction; and maternal life course. Many of the programs sought to assess results along the causal chain posited by their underlying models. Those models are illustrated in each of the articles. Results vary such that some programs have fairly consistent patterns of results, and others do not. This may be due to problems in the models or the programs, but the spotty results may also be due to the use of flawed measures to assess the outcomes.

Determining How to Measure Outcomes

Outcomes can be measured in many ways. What is easiest to document in terms of time and cost (that is, knowledge and attitude changes concerning parenting, measured through the use of paper-and-pencil ques-

tionnaires) may not be the most meaningful or the most accurate measure. For example, assessing changes in knowledge or attitudes may not be as important as assessing changes in parent-child interactions. Relying on parents' self-reports of their interactions with their children or the reports of program staff may not provide as accurate a picture of those interactions as observation by an unbiased professional not associated with the program. The more precise the measurement technique, the more dispassionate the observer, and the more policy relevant the outcome, the more costly and intrusive the evaluation is likely to be.

Evaluators prefer to rely on measures that have been tested to confirm that they are valid and reliable measures of the concepts they are supposed to assess for the population that is participating in the program. Across all the studies mentioned in this journal issue, more than 100 measures were used

The most critical choice in planning an evaluation involves whether a comparison group is included.

to assess a wide range of outcomes, and the studies reported in the main articles tended to use independent observers. Many of the measures are well known in the research literature. Not all have been investigated with the populations that used them in these studies.

Designing an Evaluation: Comparison Groups, Randomization, and Sample Size

Perhaps the most critical choice in planning an evaluation involves whether a comparison group is included. This choice determines the extent to which evaluators can reasonably claim that it was the program that caused observed benefits, because a comparison group allows a view of how families would have fared without any intervention. There are many ways to build a comparison group, with random assignment typically viewed as the best approach. Even if a study has a well-designed comparison group, if there are too few families enrolled in the

Table 1

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
CHILD OUTCOMES						
<i>Development</i>						
Catell Infant Intelligence Scale ¹	An assessment of the mental development of infants including infant verbalizations and motor control. Applicable to a younger age range than the Stanford-Binet Intelligence Scale. Examiner administered.				X	
Child Behavior Checklist (CBCL) ^{2,3}	Checklist provides profile of behavioral problems (eight or nine scales) and social competence (three scales). Provides standard scores. 100-item version for ages 2 to 3. 113-item version for ages 4 to 16 with separate norms for ages 4 to 5, 6 to 11, and 12 to 16 by gender. Parent interview.	X			X	
Bayley Scales of Infant Development (BSID and BSID-II) ⁴	A measure of infant mental (178 items) and motor (111 items) development. Used for assessing developmental progress, comparisons with peers, and eligibility for special services. Provides standard scores. For ages 2 to 30 months (BSID) or 1 to 42 months (BSID-II). Examiner administered.	X	X		X	X
Developmental Profile II (DPII) ⁵	A measure of physical, social, and mental development of children. Cognitive, communication, social, self-help, and physical development scales. For ages zero to nine. Parent and teacher interview.					X
Kaufman Assessment Battery for Children (K-ABC) ⁶	A measure of cognitive ability for ages 2 years, 6 months to 12 years, 5 months. Four scales: sequential processing, simultaneous processing, achievement, and nonverbal. Provides subset and composite standard scores. Examiner administered.	X				

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
Developmental Checklist ⁷	A 24-item measure constructed specifically for the CCDP evaluation from the Work Sampling System by Meisels. The Checklist assesses adaptive social behavior in children five years or older. Parent interview.	X				
Peabody Picture Vocabulary Test—Revised (PPVT-R) ⁸	An evaluation of the receptive vocabulary of individuals at age two years, six months through adulthood. No reading ability necessary. 175 plates with 4 pictures per plate. Provides standard scores. Also available in Spanish. Examiner administered.	X				X
Scott and Hogan Adaptive Social Behavior Inventory (ASBI) ⁹	A measure designed to assess adaptive or prosocial behaviors in high-risk three-year-olds. 30 items organized into 3 subscales: express, comply, and disrupt.	X				
Stanford-Binet Intelligence Scale—Fourth edition ¹⁰	An assessment of intelligence and cognitive abilities in verbal, abstract/visual, and quantitative reasoning and short-term memory. For ages two to adult. Provides standard scores. Administered by a certified examiner.		X		X	
School Performance						
Child Classroom Adaptation Index (CCAI) ¹¹	An 11-item rating scale measuring children's functioning and performance in the classroom. Teacher self-administered.			X		
Cooperative Preschool Inventory (CPI) ¹²	An assessment of achievement in areas necessary for school success. Used to screen children entering school and to estimate the degree of disadvantage a child may have. For preschool ages. Examiner administered.			X		

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
Metropolitan Readiness Test ¹³	An assessment of underlying skills important for early school learning. For grades K-1. Examiner administered.			X		
Metropolitan Achievement Test ¹⁴	An assessment of school achievement. Tests in reading, comprehension, mathematics, and language. For grades K-12. Examiner administered.			X		
Stanford Early School Achievement Test ¹⁵	An assessment of school achievement at the kindergarten and first-grade levels. Examiner administered.			X		
Physical Health						
Preterm birth, birth weight, gestational age	Birth outcome indicators for the child enrolled in the program. Maternal interview and child's medical records.				X	X
Access to medical care	An assessment of whether children enrolled in the program have health insurance, a regular source of preventive care, a specific primary care provider, and the ability to get care when needed. Maternal interview.		X			X
General health and health care visits	An enumeration of child visits for preventive health care and non-emergency department care. Maternal interview and child's medical records.	X	X		X	X
Immunizations	A record of appropriate childhood immunizations for age. Maternal interview and child's medical records.		X		X	X

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
Emergency department visits	A record of emergency care for treatment of illness, injuries, or ingestion of poisonous or otherwise harmful substances or objects. Maternal interview and child's medical records.		X		X	X
Hospitalizations	A record of hospital stays for any cause. Maternal interview and child's medical records.		X		X	
Abuse and/or neglect	A record of cases of child abuse and/or neglect reported to child protective services. Substantiated records are those verified by child protective services or another agency responsible for the investigation of child abuse reports.		X		X	X
Mortality	Child mortality.	X	X		X	X
PARENT OUTCOMES AND HOME ENVIRONMENT						
<i>Parenting</i>						
Adult-Adolescent Parenting Inventory (AAP) ^{16,17}	A 32-item measure of attitudes about child rearing and beliefs about parenting. For adolescents and adults. Self-administered.	X			X	
Conflict Tactics Scale (CTS2) ¹⁸	A revision of the CTS scale that assesses adult intimate partner violence to include subscales for behavior directed toward infants and children: neglect (for example, leaving alone), psychological aggression (for example, shouting), minor physical assault (for example, spanking), and severe physical assault (for example, kicking). Maternal interview.		X			
Knowledge of Infant Development Inventory (KIDI) ¹⁹	A 58-item questionnaire designed to assess knowledge of infant care, development, and behavior. Parent interview.					X

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
Home Observation for Measurement of the Environment (HOME) ^{20,21}	An assessment of the quality of the home environment for child cognitive, social, and emotional development. Two versions: infants and toddlers (45 items) and preschoolers (55 items). For ages 0 to 6. Parent and child observed behavior and parent report. Examiner administered.	X	X		X	X
Nursing Child Assessment Satellite Training (NCAST) Teaching Scale ²²⁻²⁴	An assessment of the quality of the parent-child interaction. Mother is asked to choose a task appropriate for the child's development and teach it to him/her. 73 items organized into 6 subscales. Provides standard scores. Observer rated.	X	X		X	
Parenting Sense of Competence Scale (PSOC) ^{25,26}	A 17-item questionnaire to assess attitudes about parenting and confidence in parenting ability. Parent interview or self-administered.		X			X
Parenting Stress Index (PSI) ²⁷	A measure to identify temperament and emotional problems in children and parents, and parent-child systems under stress and at risk for dysfunctional parenting. For ages 1 month to 11 years and adults. Examiner administered.		X			
Maternal Life Course						
Education attainment	Attending school, high school graduate, general educational development (GED) recipient, or postsecondary education. Maternal interview.	X	X		X	X
Marital status	Married with or without partner in the home or single with or without partner in the home. Maternal interview.	X	X		X	X

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a						
Measure and Related Endnote Numbers*	Description ^b	Program Evaluations Utilizing Measure				
		Comprehensive Child Development Program (CCDP)	Hawaii Healthy Start Program (HSP)	Home Instruction Program for Preschool Youngsters (HIPPY)	Nurse Home Visitation Program (NHVP)	Parents as Teachers (PAT)
Economic status	Job training, employment, total household income, or reliance on government benefits. Maternal interview.	X	X		X	X
Subsequent pregnancies	Number of pregnancies and births subsequent to the birth of the child enrolled in the program. Maternal interview.	X	X		X	X
Maternal Health and Behavior						
Health monitoring during pregnancy	Measures of diet, weight gain, blood pressure, sexually transmitted diseases, vaginal yeast infections, urinary tract infections, and kidney infections. Maternal interview and medical records.				X	
Substance use during pregnancy	Measures of tobacco, alcohol, and illegal drug use. Maternal interview. Urine samples (NHVP only).	X	X		X	X
Social support and connectedness	Measures of community and adult attachment and self-esteem. Maternal self-administration.	X	X			
Conflict Tactics Scale (CTS2) ¹⁸	A scale used to measure the incidence and severity of intimate partner violence. A 7-point, 39-item questionnaire to assess the extent to which dating, cohabiting, or married partners engage in psychological and physical attacks on each other and also use reasoning or negotiation to deal with conflict. Maternal interview.		X			
Mental health	Measures of psychological well-being. Maternal self-administration.	X	X		X	
Encounters with the criminal-justice system	Measures arrests, convictions, and days in jail. Maternal self-administration and criminal-justice records.				X	

* See related endnotes at the end of this table.

Table 1 (continued)

Selected Outcome Measures Used in Home Visiting Program Evaluation Studies ^a
<p>^a This table lists selected outcome measures used in the evaluations reported in this journal issue. Many of these studies used additional measures, but this table primarily describes those measures whose results are reported in the articles in this journal issue. For brevity, the more than 50 measures used in the studies of Healthy Families America (reviewed in the article by Daro and Harding in this journal issue) and of Parents as Teachers (reviewed in Appendix B in this journal issue) are not listed here.</p> <p>^b Descriptions for some of the tests presented in this table were obtained from the following: Sweetland, R.C., and Keyser, D.J. <i>Tests: A comprehensive reference for assessments in psychology, education, and business</i>. Kansas City, MO: Test Corporation of America, 1983, 1984, 1991; Maddox, T. <i>Tests: A comprehensive reference for assessments in psychology, education, and business</i>. Kansas City, MO: Test Corporation of America, 1997; and Sattler, J.M. <i>Assessment of children</i>. Third ed. San Diego, CA: Author, 1988.</p>
<p>Endnotes:</p> <p>¹ Catell, P. <i>Catell Infant Intelligence Scale</i>. San Antonio, TX: The Psychological Corporation, 1946. (NHVP-Elmira)</p> <p>² Achenbach, T.M., and Edelbrock, C. <i>Manual for the Child Behavior Checklist and Revised Behavior Checklist and Revised Behavior Profile</i>. Burlington: University of Vermont, Department of Psychiatry, 1983. (CCDP)</p> <p>³ Achenbach, T.M. <i>Child Behavior Checklist for ages 2-3</i>. Burlington: University of Vermont, 1988. (NHVP-Memphis)</p> <p>⁴ Bayley, N. <i>Bayley Scales of Infant Development</i>. San Antonio, TX: The Psychological Corporation, 1969 (CCDP;HSP;NHVP-Elmira and Memphis), 1993 (PAT).</p> <p>⁵ Alpern, G., Boll, T., and Shearer, M. <i>Developmental Profile II</i>. Los Angeles: Western Psychological Services, 1986. (PAT)</p> <p>⁶ Kaufman, A.S., and Kaufman, N.L. <i>Kaufman Assessment Battery for Children</i>. Circle Pines, MN: American Guidance Service, 1983. (CCDP)</p> <p>⁷ St.Pierre, R.G., Layzer, J.I., Goodson, B.D., and Bernstein, L.S. <i>National impact evaluation of the Comprehensive Child Development Program: Final report</i>. Cambridge, MA: Abt Associates, June 1997, pp. 5.7, 5.12. (CCDP)</p> <p>⁸ Dunn, L.M., and Dunn, L.M. <i>Peabody Picture Vocabulary Test—Revised</i>. Circle Pines, MN: American Guidance Service, 1981. (CCDP;PAT)</p> <p>⁹ Scott, K.G., and Hogan, A. <i>The Adaptive Social Behavior Inventory</i>. New York: Harcourt Brace Jovanovich, 1987. (CCDP)</p> <p>¹⁰ Thorndike, R.L., Hagen, E.P., and Sattler, J.M. <i>Stanford-Binet Intelligence Scale</i>. Fourth ed. Itasca, IL: Riverside Publishing Company, 1986. (HSP;NHVP-Elmira)</p> <p>¹¹ Halpern, R., Baker, A.J.L., and Piotrkowski, C.S. <i>The Child Classroom Adaptation Index</i>. New York: National Council of Jewish Women, 1993. (HIPPY)</p> <p>¹² Caldwell, B.M. <i>Cooperative Preschool Inventory</i>. Princeton, NJ: Educational Testing Service, 1974. (HIPPY)</p> <p>¹³ Nurss, J.R., and McGauvran, M.E. <i>Metropolitan Readiness Test</i>. San Antonio, TX: The Psychological Corporation, 1976. (HIPPY)</p> <p>¹⁴ Barlow, I.H., Farr, R., Hogan, T.P., and Prescott, G.A. <i>Metropolitan Achievement Tests: Fifth edition survey battery</i>. San Antonio, TX: The Psychological Corporation, 1978. (HIPPY)</p> <p>¹⁵ Madden, R., Gardner, E.F., and Collins, C.S. <i>Stanford Measurement Series: Stanford Early School Achievement Test</i>. Second ed. San Antonio, TX: The Psychological Corporation, 1982. (HIPPY)</p> <p>¹⁶ Bavolek, S. <i>Research and validation report of the Adult-Adolescent Parenting Inventory (AAPI)</i>. Eau Claire, WI: Family Development Resources, 1989. (CCDP)</p> <p>¹⁷ Bavolek, S. <i>Handbook for the AAPI: Adult-Adolescent Parenting Inventory</i>. Eau Claire, WI: Family Development Resources, 1984. (NHVP-Memphis)</p> <p>¹⁸ Straus, M.A., Hamby, S., Boney-McCoy, S., and Sugarman, D. <i>The revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data</i>. Durham, NH: Family Research Laboratory, University of New Hampshire, 1995. (HSP)</p> <p>¹⁹ MacPhee, D. Knowledge of Infant Development Inventory. Unpublished manuscript, University of North Carolina, Chapel Hill, 1981. Available from David L. MacPhee, Ph.D., Human Development and Family Studies, Colorado State University, Fort Collins, CO 80523. (PAT)</p> <p>²⁰ Caldwell, B.M., and Bradley, R.H. <i>Home Observation for Measurement of the Environment</i>. Little Rock: University of Arkansas, Little Rock, 1984. (CCDP;PAT, NHVP-Memphis)</p> <p>²¹ Caldwell, B., and Snyder, C. <i>Nursing Child Assessment Satellite Training/Home Observation for Measurement of the Environment</i>. Seattle, WA: NCAST. (HSP)</p> <p>²² Barnard, K. <i>NCAST Scale</i>. Seattle: University of Washington, School of Nursing, 1989. (CCDP)</p> <p>²³ Barnard, K. <i>Nursing Child Assessment Satellite Training Learning Resource Manual</i>. Seattle: University of Washington, 1987. (HSP)</p> <p>²⁴ Nursing Child Assessment Satellite Training Program. <i>NCAST Caregiver/Parent Interaction Teaching Manual</i>. Seattle: NCAST Publications, University of Washington, School of Nursing, 1994. (NHVP-Memphis)</p> <p>²⁵ Gibaud-Wallston, J., and Wandersman, L.P. Development and utility of the Parenting Sense of Competence Scale. Unpublished manuscript, Peabody College, 1976. (HSP)</p> <p>²⁶ Gibaud-Wallston, J., and Wandersman, L.P. Development and utility of the Parenting Sense of Competence Scale. Paper presented at the American Psychological Association, Toronto, Canada, 1978. (PAT)</p> <p>²⁷ Abidin, R.R. <i>Parenting Stress Index</i>. Charlottesville, VA: Institute of Clinical Psychology, University of Virginia, 1983. (HSP)</p>

program, it will be difficult to demonstrate the level of change demanded by most traditional techniques of statistical analysis.

The Need for a Comparison Group

The article by St.Pierre and Layzer concerning CCDP illustrates why comparison groups are valuable in program evaluation. The evaluators assessed the effects of the program on thousands of women and their children over the course of several years. Considering just the women and children enrolled in the program, the results suggested that the families benefitted from the services. But when families that had never been enrolled in CCDP were observed, they too had improved—about as much as the families that were enrolled in CCDP. St.Pierre and Layzer conclude that the CCDP services did not benefit the enrolled families.

Most of the evaluations in this journal issue (with the exception of some of the studies in the HFA Research Network reviewed in the article by Daro and Harding), therefore, included some sort of comparison group to make sure that the families that enrolled benefitted more than they would have without the services. The challenge for these and any evaluations is to construct the comparison group so that it is as similar to the group receiving the services as possible. If the groups are similar to begin with, and if the only way in which their experiences differ is in exposure to the home visitation program, then it is much more plausible to argue that the intervention caused any observed changes in the program group over time. Usually, comparison groups are constructed in one of three ways: (1) after the fact (post hoc), (2) at the beginning of the program through some sort of matching process (quasi-experimental), or (3) through randomized assignment (true experiments).

Post Hoc and Quasi-Experimental Designs

In some studies, the comparison group is constructed after the fact. In other words, three-year-old “graduates” from a home visitation program might be compared with three-year-olds from the community who are similar demographically (in terms of education of parents, household income, and ethnicity/race) but who have never participated in the home visitation program. This approach is used in some of the studies

that are reported in the tables for PAT (see Appendix B in this journal issue) and HFA (see the article by Daro and Harding in this journal issue). The risk in employing this approach is that the graduates may differ in some way from the comparison group, especially in motivation or drive, and those differences may account for observed program benefits. In other words, without a comparison group, it is impossible to say that the families would not have benefitted without home visiting services, because they might have sought out other opportunities on their own initiative. This is an especially serious risk in the case of home visiting programs in which attrition rates are usually quite high, suggesting that those families that remain in the programs are different from those that leave.

In other cases, the comparison group is constructed at the beginning of the program, at the same time as the intervention begins, and both intervention and comparison group participants are followed over time. Both usually receive a pretest, the intervention group receives the program services for what may be a period of years,

The challenge is to construct the comparison group so that it is as similar to the group receiving the services as possible.

and both groups take a posttest. This was the approach adopted for the Arkansas study of HIPPPY reported in the article by Baker, Piotrkowski, and Brooks-Gunn. It was also the approach used in several of the studies reported in the tables concerning PAT (see Appendix B on pages 179–89 in this journal issue) and HFA (see the article by Daro and Harding in this journal issue).

If groups are similar demographically and/or on relevant pretests, then one can argue that the home visitation program led to the observed differences between the groups at the posttest. If the groups are not equivalent initially, special statistical techniques are used to try to adjust for those differences when evaluators analyze the results.

Even if the two groups are equivalent on the pretest or statistically adjusted to look

the same, however, the groups may differ on some unobserved dimension that is the real cause of the difference. The intervention might be credited with achieving a success that was really due to that underlying factor. To combat this possibility, evaluators use a technique called random assignment to create a comparison group.

Random Assignment ("True" Experiments)

In general, random assignment involves identifying a large pool of possible program participants and then assigning them by chance (for example, by flipping a coin) to one of two or more conditions. In the simplest design, as exemplified in the CCDP evaluation, families are randomly assigned to the intervention (home visitation) group or a no-treatment control group. More complex experiments assign families randomly to one of three or four groups, which receive

Random assignment is the best way to help ensure that intervention and comparison groups are equivalent initially.

services of different types or durations. Such a design can be used to test the differential effects of a range of services. Complex designs were used in the evaluations of the Hawaii Healthy Start Program, the Nurse Home Visitation Program, and the Teen PAT study, summarized in the articles by Duggan and colleagues, by Olds and colleagues, and by Wagner and Clayton, respectively.

No matter what the variation, the rationale for random assignment remains the same: Random assignment is the best way to help ensure that intervention and comparison groups are equivalent initially. With that equivalence, evaluators can argue that results are due to the intervention and not to unmeasured differences between the groups.

Random assignment is not infallible, however, and most of the articles in this journal issue report the results of checks on whether the intervention and control groups were indeed equivalent after randomization. In some cases, they were not, and evaluators used statistical techniques

called multivariate analyses to adjust for those differences. (See the articles in this journal issue by Wagner and Clayton on PAT, by Duggan and colleagues on Hawaii's Healthy Start, and by Olds and colleagues on the Nurse Home Visitation Program.) Although such approaches are acceptable, they open the door for concern that other, unmeasured differences might exist that contribute to the observed outcomes.

Sample Size and the Rules of Statistics

Program evaluators subject the results of the evaluations to statistical tests and standardized rules to assess whether the results are truly due to the intervention and not just to chance. The tests can help determine whether the same results would be likely to recur if the study were repeated.

The general principle behind most of the statistical tests employed by evaluators is that it is harder to demonstrate a true difference between groups when the group size is small than when the group size is large. This makes intuitive sense. When groups are very small, even a fairly large percentage difference may mean a difference in only a few individuals.

For example, a decrease from 10% to 5% in child abuse rates in a group of 20 families means one child was abused instead of two. In a group of 2,000 families, the same percentage difference translates to 100 instead of 200 cases of abuse. Intuitively, evaluators can feel more confident that the 100-case difference is a real one. Many factors might account for the one-case difference in the small group, but it is harder to come up with plausible explanations other than an intervention of some kind that could have caused a difference of 100 cases of abuse.

Because of how statistical tests are calculated, statistically significant results (usually defined as results that one would expect to occur by chance no more than 5 times out of 100, and expressed as occurring with a probability of $p \leq .05$) are easier to achieve when using large groups of people than when using small groups. Program evaluators, therefore, usually want to have larger rather than smaller groups to assess.

In this journal issue, the study of the CCDP program employed the largest

groups by far (about 2,000 families each in the control and experimental groups). The other studies had smaller groups, ranging from perhaps 200 to about 40 per group for most of the main analyses and to as few as 25 or so for some of the subgroup analyses. This may mean that the groups were too small to detect some differences, especially if the differences were relatively small in magnitude.

Analyzing the Results

In the studies in this journal issue, the evaluators typically randomly assigned families to groups, and then compared the results of the groups to see if the families receiving home visiting services outperformed the families in the comparison groups. In many studies, no or only a few overall differences were seen.

For example, in the Elmira study of nurse home visitation, when comparing all the participants in the home visitation group with all the participants in the comparison group, Olds and colleagues found no differences in measures such as child development or birth outcomes. However, striking differences were found when the analyses were restricted to unmarried, low-income women. For some analyses, Olds and colleagues compared the unmarried, low-income women in the home visitation group with the unmarried, low-income women in the control group and found that home visitation for those families had indeed produced benefits in several domains, including deferral of subsequent pregnancies.

In their report on PAT, Wagner and Clayton employ similar subgroup analyses to assess the differential impact of the program on Caucasian, English-speaking Latino, and Spanish-speaking Latino families. Other articles also report differences among subgroups of the participants.

The risk of this approach is that such analyses can essentially eliminate the benefits of randomization—unless the evaluators plan ahead for these sorts of analyses by randomly assigning families to groups using special techniques to stratify the groups initially. This means that the evaluators would randomize the groups such that the same number of families with each particular characteristic being analyzed would wind up

in each group (for example, the same number of unmarried women, English-speaking Latinos, and so on).

Without this initial step, it is possible that the families in the intervention and control subgroups will somehow differ from one another in unmeasured ways, and that any differences that are found will really be a reflection of those underlying differences, and not of their exposure to home visitation. Researchers often use statistical techniques to try to equalize the groups, but there is no guarantee that the techniques will be successful.

The most conservative approach, therefore, is to treat subgroup analyses that were not part of the initial analytic plan for an evaluation as preliminary findings. It is quite plausible that a given home visiting program will be more effective with some

Researchers should seek to build upon the subgroup findings from one study to see if they appear in subsequent, more carefully controlled studies.

families than with others, but researchers should seek to build upon the subgroup findings from one study to see if they appear in subsequent, more carefully controlled studies. Olds and colleagues report on a line of research that follows this approach.

Interpreting the Results

Methodologically strong evaluations with sensible analytic plans still require interpretation. Factors that may influence how much weight to give to a single evaluation include attrition from the study, the policy and practical relevance of its findings, and the likelihood that its results are generalizable across multiple settings.

Attrition

The articles in this journal issue report that between 20% and 67% of families leave the home visiting programs before their intended completion. Families leave for a variety of reasons, including mobility out of the community, a lack of interest, and, perhaps, a belief that they have already

derived as much benefit as they can from the programs.

Such attrition may indicate that program modification is needed to increase family engagement, but it is also a sign that the evaluation could be weakened. If those who remain in the program are somehow different from those who have dropped out (perhaps because they are more motivated to seek improvement), then an evaluation that assesses only those families that remain in the program may overestimate the program's benefits. Those persevering families might well have benefitted without home visiting services, because they might have sought out other community services or resources on their own.

Most methodologists believe that the most appropriate way to assess a program in the face of attrition is by measuring all the families, whether or not they receive the intended services, keeping them in their

ment is not the same as winning the war of policy or practical relevance. Results should be examined to determine whether the questions they investigated are still timely. Since the studies of HIPPI and PAT reported in this journal issue were mounted, for instance, the service models have evolved, as reported in Appendix D on pages 192–94, and Appendix B on pages 179–89, in this journal issue, and it is not clear how many program sites are operating the older variations of the model rather than the newer ones.

Policymakers and practitioners should also consider the functional importance of a program's results. A one- or two-point difference between huge groups on a paper-and-pencil test of parent attitudes or knowledge may be statistically significant, but it may have little or no importance for public policy or practice because the connection between results on that test and the outcome of ultimate interest (rates of child abuse and neglect) is too tenuous.

Similarly, a decrease of 1% or 2% in teen pregnancy rates will not be nearly as persuasive to a policymaker as a 10% shift. Indeed, policymakers may not judge a program a success unless it generates effects of a particular size, often because it is only when effects are large enough in magnitude that they produce cost savings.

In their review of CCDP, St. Pierre and Layzer note a few outcomes in which differences were statistically significant but too small, in their opinion, to be meaningful in terms of children's development. None of the other researchers used this approach, and policymakers and practitioners will undoubtedly bring their own lens to these studies.

Generalizability

A rigorous evaluation, unmarred by high attrition and revealing important outcomes, can help demonstrate that a program has worked in a particular setting. Unfortunately, no single evaluation can demonstrate that the program will work equally well in another setting. The abilities of administrators and home visitors will differ across agencies, and the needs of families will vary across communities. Ancillary community services may be of high quality

Policymakers and practitioners should consider the functional importance of a program's results.

originally assigned groups.¹ This is often called the intention-to-treat approach. It is expensive because it requires that evaluators locate families that have moved away from the area, and that they pursue individuals who may prefer to be left alone.

This approach was attempted in most of the studies reported on in this journal issue, but not all of the studies were able to locate all of the individuals who disappeared from the program. Attrition from the evaluations ranged from 11% to 48%, hovering around 20% for most. In most cases, the program evaluators sought to demonstrate that the smaller groups were initially equivalent to the larger groups in terms of background characteristics such as age, ethnicity, income, and education, but there is no way to tell whether the groups differed on intangibles such as motivation or drive.

Statistical Significance Versus Policy Relevance

Winning the battles of statistical significance, research rigor, and family engage-

in one community but not in another. The program itself may be modified in different settings—by intention, to better serve the needs of a new population; by accident on the part of new practitioners not familiar with the model; or by necessity, to keep costs down. The article by Duggan and colleagues reviewing Hawaii's Healthy Start Program details the ways in which different agencies, all implementing the same model, operate it differently, with dramatic differences in client participation and outcomes.

All of the articles in this journal issue report the results of studies of a model at more than one site. The variability of results across sites suggests that generalizability of results may be limited.

Summary

Most of the major evaluations summarized in the articles in this journal issue are of reasonably good quality: all include comparison groups, most of them developed through randomized assignment; many include mea-

asures of both implementation and outcome; and many assess a variety of outcomes. However, they have weaknesses, too. The individualization of service content and delivery inherent in home visiting programs may make it hard to see differences across a whole group, because, in fact, the group is not getting the same treatment. A mix of standardized tests and less-well-recognized measures, not all of which had been previously tested with similar populations, were used. The relatively high attrition rates from some programs, and the lower, but sometimes still high attrition rates from their studies may weaken some of the conclusions that can be drawn from the evaluations. Results vary across program sites, suggesting that generalization of results may be limited.

Nevertheless, these evaluations stand as some of the best in the home visiting field. Interpretation of evaluation results is an art, not a science, but these studies are sturdy enough to provide guidance to policymakers and practitioners.

1. McKinlay, S.J., Stone, E.J., and Zucker, D.M. Research design and analysis issues. *Health Education Quarterly* (Summer 1989) 16:307–13.

A number of excellent references on program evaluation are available:

- Campbell, D.T., and Stanley, J.C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1963.
- Cook, T.D., and Campbell, D.T. *Quasi-experimentation: Design and analysis issues for field settings*. New York: Rand-McNally, 1979.
- Cronbach, L., Ambron, S., Dornbusch, S., et al. *Toward reform of program evaluation*. San Francisco: Jossey-Bass, 1980.
- Pawl, J., Barnard, K., Korner, A., et al. *Charting change in infants, families, and services: A guide to program evaluation for administrators and practitioners*. Washington, DC: National Center for Clinical Infant Programs, 1987.
- Rossi, P.H., and Freeman, H.E. *Evaluation: A systematic approach*. Fourth ed. New York: Sage Publications, 1989.
- Weiss, H.B., and Jacobs, F.H. *Evaluating family programs*. Hawthorne, NY: Aldine de Gruyter, 1988.