Lessons Learned from the Home Visiting Evidence of Effectiveness Review

January 2011

Contract Number:
HHSP23320095642WC/HHSP23337007T

Mathematica Reference Number:
06686-610

Submitted to:
Lauren Supplee, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Project Director: Diane Paulsell
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005

Lessons Learned from the Home Visiting Evidence of Effectiveness Review

January 31, 2011

Sarah Avellar
Diane Paulsell

This report is in the public domain. Permission to reproduce is not necessary. Suggested citation: Avellar, S., and Paulsell, D. (2011). *Lessons Learned from the Home Visiting Evidence of Effectiveness Review* Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at http://www.acf.hhs.gov/programs/opre/index.html.

MATHEMATICA
Policy Research, Inc.

LESSONS LEARNED FROM THE HOME VISITING EVIDENCE OF EFFECTIVENESS REVIEW

Early childhood home visiting is a strategy for delivering a broad range of services and supports to families with pregnant women and young children, primarily during visits to families' homes. Home visiting to provide public health and early education services dates back to the nineteenth century; since the 1960s, this strategy has increased in prominence as more national home visiting models have been developed and tested (Wasik & Bryant, 2001; Boller, Strong, & Daro, 2010; Stoltzfus & Lynch, 2009). Especially in the past 30 years, researchers have rigorously evaluated a growing number of program models and have identified promising outcomes in a range of domains, including maternal and child health, child development, parenting, and family economic self-sufficiency (Daro, 2006; Gomby, 2005). Early childhood home visiting programs currently operate in all 50 states, with an estimated 400,000 to 500,000 families receiving services (Stolzfus & Lynch, 2009). Moreover, according to a recent study conducted by the Pew Center on the States, 46 states operate at least one state-administered early childhood home visiting program (Pew Center on the States, 2010).

The Patient Protection and Affordable Care Act significantly expands federal funding for home visiting by establishing the Maternal, Infant, and Early Childhood Home Visiting Program (MIECHV). This new initiative is providing $1.5 billion to states over five years to establish home visiting programs for at-risk pregnant women and children from birth to age 5. The act stipulates that 75 percent of the funds must be used for home visiting programs with evidence of effectiveness based on rigorous evaluation research. In preparation for this new initiative, the Office of Planning, Research, and Evaluation at the Administration for Children and Families/Department of Health and Human Services (DHHS), contracted with Mathematica Policy Research in fall 2009 to launch Home Visiting Evidence of Effectiveness (HomVEE), a systematic review of home visiting research.

HomVEE conducted a review of the home visiting research literature and assessed the evidence of effectiveness of home visiting program models that serve families with pregnant women and children from birth to age 5. The HomVEE review provides states and other stakeholders with information about which home visiting models have evidence of effectiveness as required by the legislation as well as with detailed information about the samples of families who participated in the research, the outcomes measured in each study, and the implementation guidelines for each model. A summary of the review findings is available in the HomVEE Executive Summary (Paulsell, Avellar, Sama Martin, & Del Grosso, 2010); detailed findings are available on the HomVEE website (http://www.acf.hhs.gov/programs/opre/homvee).

This paper describes key lessons learned from the first year of the HomVEE review about the current state of evidence on the effectiveness of early childhood home visiting, gaps in the research literature that create challenges for assessing effectiveness, and suggestions for strengthening future research in this area. The first section provides a brief overview of the review process and standards. Section II discusses the range of research designs used in the studies reviewed and issues related to assessing replication of findings. Section III addresses measurement issues, and Section IV describes challenges related to determining whether effects are meaningful in terms of their magnitude, breadth, consistency, and duration. This section also discusses interpretation of subgroup findings. In sections V and VI, the paper discusses implementation issues that are important for assessing effectiveness, including fidelity of implementation, adaptation of program models, and the context for implementation and evaluation. Section VII discusses conflict of interest issues of relevance to the home visiting research literature. Sections VIII and IX provide a summary of key lessons learned about the home visiting research literature and offer suggestions for strengthening future research.

I. HomVEE Review Process, Standards, and Findings

The HomVEE team carried out seven main activities to conduct a systematic review of the home visiting research literature: (1) conducted a broad literature search, (2) screened studies for relevance, (3) prioritized program models for review, (4) rated the quality of impact studies, (5) assessed the evidence of effectiveness for each program model, (6) reviewed implementation information for each program model, and (7) addressed conflicts of interest.

Literature Search and Screening Process. HomVEE limited the literature search to research on program models that used home visiting as the primary service delivery strategy, offered home visits to most or all participants, and aimed to improve outcomes in at least one of eight domains specified in the legislation: (1) child development and school readiness; (2) child health; (3) family economic self-sufficiency; (4) linkages and referrals; (5) maternal health; (6) positive parenting practices; (7) reductions in child maltreatment; and (8) reductions in juvenile delinquency, family violence, and crime. The literature search included a database search, website searches, a call for studies, and an examination of existing literature reviews and meta-analyses. HomVEE then screened out studies that were not relevant and ranked program models for review based on the number and design of impact studies, sample sizes of impact studies, availability of implementation information, and prevalence in the field.

This process yielded 11 prioritized home visiting models for review:

1. Early Head Start-Home Visiting
2. Family Check-Up
3. Healthy Families America (HFA)
4. Healthy Start-Home Visiting
5. Healthy Steps
6. Home Instruction for Parents of Preschool Youngsters (HIPPY)
7. Nurse Family Partnership (NFP)
8. Parent-Child Home Program
9. Parents as Teachers (PAT)
10. Resource Mothers Program
11. SafeCare

Rating the Quality of Impact Studies. HomVEE reviewed impact studies of the prioritized models with two types of designs: randomized controlled trials and quasi-experimental designs[1]

[1] HomVEE defines a quasi-experimental design as a study design in which sample members (children, parents, or families) are selected for the program and comparison conditions in a nonrandom way.

(including matched comparison group designs, single case designs, and regression discontinuity designs). Trained reviewers assessed the research design and methodology of each study using a standard review protocol. Each study was assigned a rating of high, moderate, or low to provide an indication of the study design's capacity to provide unbiased estimates of program impacts.

In brief, the high rating was reserved for random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment as well as single case and regression discontinuity designs that meet What Works Clearinghouse (WWC) design standards.[2] The moderate rating applied to random assignment studies that, due to flaws in the study design, execution, or analysis (for example, high sample attrition), did not meet all the criteria for the high rating; matched comparison group designs that established baseline equivalence on selected measures; and single case and regression discontinuity designs that met WWC design standards with reservations. Studies that did not meet all of the criteria for either the high or moderate rating were assigned the low rating.

Assessing Evidence of Effectiveness. After completing all impact study reviews for a program model, HomVEE evaluated the evidence across all studies of the model that received a high or moderate rating. To meet the HHS criterion for an "evidence-based early childhood home visiting service delivery model," models must meet at least one of the following criteria:

- At least one high- or moderate-quality impact study of the model finds favorable, statistically significant impacts in two or more of the eight outcome domains.
- At least two high- or moderate-quality impact studies of the model using non-overlapping analytic study samples find one or more favorable, statistically significant impacts in the same domain.

In both cases, the impacts considered must either (1) be found for the full sample or (2) if found for subgroups but not for the full sample, be replicated in the same domain in two or more studies using non-overlapping analytic study samples. Additionally, following the legislation, if the model meets the above criteria based on findings from randomized controlled trial(s) only, then one or more favorable, statistically significant impacts must be sustained for at least one year after program enrollment, and one or more favorable, statistically significant impacts must be reported in a peer-reviewed journal.[3]

Review Results. Seven of the eleven prioritized models met these criteria: (1) Early Head Start-Home Visiting, (2) Family Check-Up, (3) HFA, (4) Healthy Steps, (5) HIPPY, (6) NFP, and (7) PAT. In addition to assessing whether each model met the HHS criteria for an evidence-based early childhood home visiting service delivery model, the HomVEE team examined and reported on other aspects of the evidence for each program model, including the quality of the outcome measures, duration of the impacts, replication of the impacts, subgroup findings, unfavorable or ambiguous impacts, and evaluator independence. For each prioritized program model, the HomVEE team also developed detailed implementation profiles based on reviews of

[2] The What Works Clearinghouse (WWC), established by the Institute for Education Sciences in the U.S. Department of Education, reviews education research.

[3] Section 511 (d)(3)(A)(i)(II)(aa).

implementation studies, impact studies with a high or moderate rating, and other implementation materials and guidance available from program developers and national program offices. Finally, HomVEE ensured that all team members disclosed any financial or personal connections to developers, studies, or products being reviewed. Team members with potential conflicts were excluded from reviews related to those programs.[4]

II. Research Design: Linking Cause and Effect

Assessing whether a program is effective requires a study design that can establish that a program *caused* the observed outcomes. A study's potential to establish causality and rule out other reasons for the observed outcomes is known as internal validity. To link a program and outcomes, a study tries to establish the counterfactual: what would have happened in absence of the program. The ideal—and impossible—method for determining the counterfactual is to observe the same group simultaneously receiving and not receiving the program. Without the possibility of establishing the true counterfactual, study designs with strong internal validity use a comparison group or condition, which is intended to represent what would have happened to the treatment group in the absence of the program.

The HomVEE review rated studies on their ability to produce unbiased estimates of a program's effect, which requires strong internal validity. The rating system helps distinguish between studies in which we have more confidence that the observed findings were caused by the program and studies in which the observed findings may be the result of other unobserved differences between the program and comparison conditions, such as participant motivation. Only studies where the selection process for these conditions is completely known, including in a randomized controlled trial, single case design, and regression discontinuity design, can receive the highest rating.

Many program models had at least one study using a randomized controlled trial, a design with strong internal validity.

A randomized controlled trial (RCT)—where participants are assigned to the treatment or comparison groups by chance—has the potential for strong internal validity. The primary advantage of randomly assigning participants is that the groups are balanced, on average, for characteristics that are known, such as race and ethnicity and education, and characteristics that may be unknown, such as patience or motivation. If the groups are the same before the program, any post-treatment differences between the groups that are too large to be due to chance are attributable to the program. However, certain factors—such as the number of participants who drop out of the study—can compromise the design and weaken the study's ability to draw causal conclusions. In the HomVEE review, an RCT could receive a high, moderate, or low study-quality rating depending on the presence of these factors.

Of the 11 prioritized models, 7 had at least one RCT. Not all of those studies received the highest study-quality rating because some suffered from high attrition, confounding factors, or other issues. However, six of the prioritized models had at least one study with the highest rating. By conducting RCTs, which can be expensive and difficult to implement, the home visiting field shows

[4] Contracted reviewers who were not Mathematica employees reviewed studies of programs previously evaluated by Mathematica Policy Research.

an understanding of the importance of using rigorous research methods to assess program effectiveness.

The review identified few single case or regression discontinuity designs.

In a single case design (SCD), the same case, which can be an individual or group, serves as its own control. This differs from a pre/post design, however, because multiple measures of the outcome are taken before and after the program. Thus, a trend of performance can be established prior to, during, and/or after the program. Further, the demonstration of an effect can be replicated in various ways, for example, if a program is introduced at different times to different families (sometimes called a multiple baseline design). The effect is replicated if it is shown for each family after the introduction of the program, regardless of the timing of the program's introduction. To receive a high rating in the HomVEE review, a study had to include at least three attempts to demonstrate an effect, systematically manipulate the introduction and withdrawal of a program, establish inter-assessor agreement on the outcomes, and have at least five data points in each phase.

In the HomVEE review, we identified two SCD studies conducted on one of the prioritized models. Both of these studies failed to meet the criteria for a high or moderate rating. Thus although SCDs have the potential for establishing the effectiveness of programs, so far the research has been of low quality, according to HomVEE standards.

Regression discontinuity (RD) is another design that can establish a strong causal link between a program and outcomes. In an RD design, the sample is assigned to treatment and comparison conditions based on the value of a "scoring" variable. For example, a home visiting program might be offered to families with infants who were born below a certain birth weight and not offered to families with infants born at or above that weight. Because the selection process is known and can be perfectly measured—unlike quasi-experimental designs with comparison groups formed in some other way—the analysis can adjust for differences in selection to produce an unbiased estimate (Shadish, Cook, & Campbell, 2002). To receive the high rating, a study must meet certain criteria, such as maintaining the integrity of the scoring variable (that is, no manipulation of the selection process), meeting WWC standards for attrition, and using an appropriate analysis. Studies that did not meet these criteria received a moderate or low rating.

In the initial review of the 11 prioritized models, HomVEE did not identify any studies that used an RD design. This design may be gaining in popularity, as researchers learn of its advantages, but in the home visiting field, it has not been widely used.

Conducting a strong study using a quasi-experimental design can be challenging. None of the matched comparison group quasi-experimental designs met HomVEE standards.

Matched comparison quasi-experimental designs (QEDs), which use a nonrandom process for group assignment, could have received a moderate study-quality rating in the HomVEE review. This purposeful selection process can compromise the quality of the QED. If the groups are different at onset, the comparison group does not provide a good representation of what would have happened to the treatment group without the program. The HomVEE review standards required that QEDs establish baseline equivalence between the two groups on selected measures. These measures, such as pre-program outcomes, race, ethnicity, and socioeconomic status, were determined to be key for composing a reasonable comparison. Regardless of how balanced the treatment and comparison groups are on measured characteristics, however, the weakness of a quasi-experimental design with a comparison group is that it can never rule out differences in unmeasured characteristics. Therefore, the conclusions from a quasi-experimental design are suggestive of an initiative's effectiveness but cannot definitely determine causality.

In the HomVEE review, none of the QEDs with matched comparison groups received a moderate quality rating; all received a low rating. Many of the studies received a low rating because (1) the treatment and comparison groups differed on key baseline characteristics or (2) information on baseline characteristics was not presented and equivalence could not be determined. Without evidence of baseline equivalence, we cannot determine how well the comparison group represents what would have happened to the treatment group in absence of the program.

Few home visiting models have been rigorously tested through replication. The number of publications does not necessarily represent the number of separate studies.

Replication is an important component of determining a program model's effectiveness. Small, convenience samples were used in many studies of home visiting models, and results may not generalize to other groups. In addition, developers often are involved in early evaluations, and some offer the services directly, which may not represent a "typical" provider. Replication increases confidence that findings were not due to chance in an isolated study (Society for Prevention Research, n.d.)

Several of the models had a large number of publications identified in the literature search, but typically were not replications because the studies did not always represent separate analytic samples or studies. For example, NFP had 16 studies that met HomVEE standards for high or moderate quality. The majority of those studies were based on RCTs conducted in Elmira, Memphis, and Denver. Similarly, HFA had numerous publications on trials conducted in Hawaii, Alaska, and New York. Separate publications are likely when follow-ups are conducted and results reported over time or when subgroups are analyzed.

In Table 1, we indicate the number of RCTs conducted for each of the 11 prioritized models, the only design in the HomVEE review that received a high or moderate rating. We do not include other designs because all reviewed studies that used other designs received a low rating. These results are based on the number of separate samples that were randomly assigned and include RCTs with both high or moderate study-quality ratings.

Although many of the prioritized models had multiple RCTs, few had true replications. To be considered a true replication, studies should include samples with the same characteristics and use the same outcomes. Ideally, a replication is conducted by a separate independent research team. Replications of results shows that findings were not limited to one particular study or sample, and thus increases confidence in the findings (Society for Prevention Research, n.d.). Separate studies are legitimate, however, and may demonstrate whether the program model is effective in other circumstances. But in the HomVEE review, because of the use of different outcomes or samples, typically, later studies could not be used to confirm earlier results.

Table 1. Number of Randomized Controlled Trials by Model

Program	Number of RCTs Conducted
Early Head Start-Home Visiting	1
Family Check-Up	2
Healthy Families America (HFA)	5
Healthy Start-Home Visiting	0
Healthy Steps	1
Home Instruction Program for Preschool Youngsters (HIPPY)	2
Nurse Family Partnership (NFP)	4
Parent-Child Home Program	0
Parents as Teachers (PAT)	4
Resource Mothers Program	0
SafeCare	0

III. Outcomes: Accurately Measuring the Targets of Change

In addition to a rigorous study design that can provide unbiased impact estimates, studies must accurately measure child and family outcomes that are the targets of change. In this section we discuss outcome measurement in the home visiting research literature, including the quality of measures used and the large number of outcomes assessed in many studies.

Studies used a wide range of measures across domains, and measures vary in quality.

Studies in the home visiting literature tend to measure outcomes in multiple outcome domains and use a wide variety of measures to do so. Many studies relied on administrative records; direct assessments and observations; and well-known, standardized parent report measures. Studies also reported on the properties of the measures, such as their internal consistency and inter-rater reliability. Others relied more on nonstandardized self-reports and did not consistently report the properties of the measures. In general, the wide range of measures creates challenges for comparing results across studies.

HomVEE sought to differentiate outcome measures based on the likelihood that they accurately measure the outcome of interest. Measures should be reliable, producing similar results with the same level of accuracy each time they are administered, and valid, accurately representing the construct of interest. To distinguish between outcome measures with potential to yield results at different levels of accuracy, the HomVEE review categorized measures used in the studies reviewed as primary or secondary (Table 2).

HomVEE has more confidence in primary measures, which include direct assessments; direct observations; data extracted from medical, school, or administrative records; and parent and teacher reports based on standardized measures. When a measure is standardized, it is administered using a uniform set of procedures for administration and scoring and uses established scoring norms based on the performance of a norming sample. Secondary measures are nonstandardized parent, teacher, or youth self-reports. For example, HomVEE classified counts of health care services received based on medical records as primary and parent reports of health care services (such as number of doctor visits or child immunizations) as secondary. Likewise, HomVEE classified the Home Observation for Measurement of the Environment (HOME; Caldwell & Bradley, 1984), a direct observation measure, as a primary outcome measure in the parenting domain and classified parent self-reports on parenting attitudes and skills based on nonstandardized measures as a secondary outcome measure.

Table 2. Primary and Secondary Outcome Measures Reported in the HomVEE Review

Outcome Domain	Primary Measures	Secondary Measures
Child health	- Birth outcomes and counts of health care service use from medical records	- Parent reports of child health status and use of health care services
Child development and school readiness	- Direct child assessments - Direct observations of children's behavior - Parent and teacher reports on standardized measures	- Parent and teacher reports on nonstandardized measures
Family economic self-sufficiency	- Administrative records of public assistance receipt	- Parent reports of public assistance receipt, employment, and economic outcomes
Linkages and referrals	- Referrals documented in home visitor, medical, or school records	- Parent reports of referral receipt and awareness of other services available in the community
Maternal health	- Counts of health care service use from medical records - Maternal reports on standardized measures	- Maternal reports on nonstandardized measures
Positive parenting practices	- Observations of parent-child interactions via live observation or video recording - Observations of the home environment - Parent reports of parenting attitudes and practices using standardized measures	- Parent reports of parenting attitudes and practices using nonstandardized measures
Reductions in child maltreatment	- Substantiated reports of child maltreatment from administrative records - Medical records of health care service use for injuries or ingestions	- Parent reports of health care service use for injuries or ingestions - Conflict Tactics Scale-Parent Child
Reductions in juvenile delinquency, family violence, and crime	- Incidents of parent or youth antisocial behavior based on administrative records	- Parent, teacher, and youth self-report of antisocial behaviors - Conflict Tactics Scale

Many studies examine a large number of outcomes but do not correct for multiple comparisons.

Many home visiting models strive to change numerous facets of families' well-being and thus measure outcomes in multiple domains to test the effectiveness of the program. As noted earlier, the HomVEE review included eight outcome domains, and most of the prioritized program models that had at least one high or moderate study-quality rating measured multiple outcomes within and across domains. The number of tested outcomes is further increased through multiple follow-ups, which were used to determine if effects were sustained over time or if new effects manifested over time. For example, the same outcome may be tested in a sample of families at 6 months, 1 year, 2 years, and so on after program enrollment.

In the HomVEE review, we selected a cutoff, or alpha level, of 0.05 for statistical significance, which means that there is a five percent chance or less of finding a false positive for one outcome. The probability of a false positive, however, increases when multiple outcomes are tested. For example, if separate *t*-tests are conducted, each with an alpha level of 0.05, then when there are five outcomes tested, the probability of a false positive is now 23 percent, and when 20 outcomes are tested, the probability rises to 64 percent (Schochet, 2009).

To counteract this increase in the risk of a false positive, corrections can be made. For example, one correction, known as the Bonferroni procedure, divides the alpha level by the number of outcomes so that the total probability across all outcomes is 5 percent. For example, if 10 outcomes are tested, the cutoff for each outcome is 0.005 for statistical significance. Other, less conservative, corrections are also available (see Schochet, 2009, for a discussion).

Very few of the studies reviewed for HomVEE made corrections for multiple comparisons, even when multiple tests were conducted. Further, HomVEE was not able to apply the correction because of missing information in many studies, and it would have been inconsistent to apply the corrections for some studies and not others. Therefore the statistical significance of these findings should be interpreted with caution.

IV. Effectiveness: Detecting Meaningful Effects

The effectiveness of a program is determined by treatment and comparison differences, with the expectation that the outcomes of those in the treatment group improved relative to the comparison group. However, the question of what constitutes a meaningful difference can be a thorny issue. For example, how large does a difference have to be to have a meaningful difference for children and families? How long does an effect have to be sustained? Is statistical significance sufficient for determining whether an effect is meaningful? Are the effects similar for key subgroups?

When possible, HomVEE reported effect sizes, which provide a measurement of the strength of the relationship.

An effect size shows the size of the impact (or the difference between the program and comparison group) relative to the standard deviation of the measure. Effect sizes can be useful for comparing or summarizing results from different studies (Hedges, 2008). Since effect sizes are expressed relative to the standard deviation, they are independent of the units of measurement for the outcome, and are particularly useful when studies have not used the same outcomes. Unlike statistical significance, the effect size is not affected by the sample size but represents the size or strength of the relationship without regard to how precisely the size of that relationship was estimated (Hedges, 2008).

Although effect sizes provide information on the strength of the relationship, the clinical or practical significance is not always the same. A commonly cited guideline is that "small" effects are about 0.20 to 0.30, "moderate" effects range from about 0.30 to 0.50, and "large" effects are greater than 0.50 or half of a standard deviation (Cohen, 1988; Lipsey, 1990). The implications for effect sizes, however, could differ across domains. For example, a small effect size in the domain of child maltreatment or juvenile delinquency may have great practical, clinical, and cost consequences, whereas a larger effect size in maternal health may have fewer implications for the well-being of the family. Similarly, a small change early in a child's life could change that child's trajectory and have long-lasting implications whereas a small change later may not result in so many cumulative changes.

When possible, the HomVEE review reported effect sizes, either those calculated by the author or HomVEE computed findings, but DHHS did not factor the size of the effect into its criteria. Instead, the HomVEE standards considered the statistical significance of the effects. One reason the standards rely on statistical significance rather than effect sizes is that for some outcomes an effect size was not available. It was not reported by the authors and HomVEE did not have sufficient information from the article to calculate it. Without comprehensive information, this would require making judgments on some effects and not others. In addition, there are general guidelines on the magnitude of an effect, as described above, but the meaning could differ across domains.

Many of the models showed longer-term, sustained findings.

Many of the program models reviewed by HomVEE were expected to have long-term results. For example, NFP states that the model "fosters long-term success for first-time moms, their babies, and society" (NFP, 2010). Similarly, the stated mission of PAT is "to provide the information, support and encouragement parents need to help their children develop optimally during the crucial early years of life" (PAT, 2010).

Of the 11 prioritized models, 7 were shown to have sustained findings for at least one year after program enrollment, the DHHS criterion. However, we often found that favorable, statistically significant findings in shorter-term follow-ups were not the same outcomes that were significant in later follow ups. It is difficult to determine why this occurs, whether some effects fade out over time or whether there are "sleeper" effects, when a finding becomes statistically significant over time. This could happen if an intervening variable was changed, leading to a subsequent effect, such as increased parent book reading eventually leading to improved child language scores.

The HomVEE review also examined outcomes measured one year after program completion. This measure provides an indication of program effects after the program has ended. However, most studies of the prioritized program models only included outcomes measured during program implementation or immediately after program completion. Thus for most programs, HomVEE was not able to assess whether any program effects were sustained over time after the program ended.

The DHHS criteria include an indicator of the consistency of favorable findings, which is supplemented by additional information in the HomVEE reports.

The DHHS criteria include an indicator of consistency because a program model ideally exhibits consistent findings for favorable impacts. The DHHS criterion requires models to show favorable, statistically significant effects in two domains within a study or to show two effects in the same domain across two different analytic samples.

Seven of the 11 prioritized models were able to meet the DHHS criterion for consistency, in conjunction with the other requirements described earlier. However, we often found that similar findings within domains were not consistently statistically significant. For example, a measure of child health that was favorable during one follow-up was not statistically significant in later follow-ups. In addition, for all models with a high or moderate quality study, we found that most outcomes measured within a domain were not statistically significant. The number of findings that were not statistically significant typically was greater than the number of findings that were statistically significant and favorable. Thus, the favorable or unfavorable findings were typically not part of a larger pattern of statistically significant findings.

Findings, categorized as favorable, no effect, or unfavorable/ambiguous, are listed on the first page of the HomVEE model reports, so readers can see the patterns. Questions to consider when evaluating these results include the following:

- Is there a pattern of favorable results among similar outcomes?
- Are there any unfavorable or ambiguous results?
- Are there any results that show no detectable effect?
- If the results are measured over time, are there consistent patterns of favorable, unfavorable, or no effect findings on the same outcomes?

There is little information on the effectiveness of program models for certain populations, including different races and ethnicities and military families. Some groups, such as Native Americans and military families, were included in samples in very small numbers, if at all.

Although several programs have at least one high-quality study, there is little evidence of effectiveness among important subgroups, such as African Americans, Hispanics, Native Americans, recent immigrants, or military families. Most studies were conducted with samples that included multiple racial and ethnic compositions, but the results were generally aggregated across groups. Thus the available evidence does not indicate whether a model is more (or less) effective with one group or another. Further, for groups that are generally not included in samples or included only in very small numbers, such as Native Americans and military families, the aggregated findings will be dominated by families with other characteristics.

Despite the potential utility of subgroup findings, such findings may need to be interpreted with caution. Subgroup findings can be identified with data mining, that is, selecting subgroups based on available data rather than based on theoretically grounded definitions (for example, a model includes adaptations for a particular population). The risk of data mining is described above: multiple comparisons increase the likelihood of finding a statistically significant finding by chance. To circumvent this problem, subgroups of interest should be determined prior to any data analysis using characteristics that are measured at baseline and thus could not be affected by the intervention. Conversely, the results for subgroups may be underpowered—that is, findings may not be shown to be statistically significant even if there is a true association—because of small sample sizes. For a given sample size, the power to detect effects for subgroups is linked with the size of the subgroup: as the proportion of the subgroup decreases so does the power (Brown 2009; Klerman 2009). In fact, appropriately powering subgroup analyses requires much larger samples than those found in the studies reviewed by HomVEE and may not be feasible given the added cost.

Balancing the interest in subgroups with the potential challenges, HomVEE only reported subgroup results for findings that were replicated in two different analytic samples. A replication of results provides greater confidence that the subgroup findings were not observed by chance. Most subgroup findings in the studies reviewed did not meet HomVEE's replication requirements. Subgroup findings were replicated and reported for NFP only. NFP reported results for mothers with low psychological resources in at least two separate samples.

V. Variations in Implementation: Fidelity to the Program Model

The best test of effectiveness of an intervention occurs when the program model is implemented with a high degree of fidelity to the original program design. This ensures that the program being evaluated was actually implemented as intended by the developer (Dane & Schneider, 1998; O'Donnell, 2008). Although consensus on a single definition does not exist, five elements are common to many definitions of implementation fidelity: (1) adherence to the program model as described by the developer, (2) exposure or dosage, (3) quality of program service delivery, (4) participant responsiveness, and (5) understanding of the essential program elements that cannot be subject to adaptation (Dunsenbury, Brannigan, Falco, & Hanson, 2003; Carroll et al., 2007).

In the home visiting research literature, several important issues related to fidelity were identified: measurement of fidelity in program evaluations, evolution of the program model over time, and adaptations to the program model.

HomVEE did not consider fidelity of implementation in program ratings because most of the causal studies reviewed did not include assessments of fidelity.

Several program models included in the HomVEE review have fidelity standards and measures that are to be used by programs and national support offices to monitor programs' fidelity of implementation. Few of the causal studies included in the HomVEE review, however, included assessments of fidelity in their evaluations. Without measuring fidelity, studies run the risk of drawing incorrect conclusions that a program model is not effective when in fact it is effective but was not implemented correctly, or that an intervention is effective when in fact the model was not implemented according to the developer's specifications (Knoche, Sheridan, Edwards, & Osborn, 2010). This problem is sometimes referred to as a Type III error, in which the research questions and the focus of the research are not aligned (Dobson & Cooke, 1980; Scanlon, Horst, Nay, Schmidt, & Waller, 1979; Knoche et al., 2010).

However, the converse problem is using only studies or samples that have strong fidelity to show evidence of effects. It is problematic to "cherry pick" samples because of potential selection bias, such as if more dedicated home visitors are more likely to implement the program with fidelity. Developer involvement to ensure fidelity is common in early efficacy trials. Later effectiveness trials, in which fidelity is likely to vary across home visitors and sites, can provide important information about levels of fidelity likely to be obtained in real-world settings in which developers are not directly involved, and the effects on child and family outcomes real-world implementation can produce. Overall, studies reviewed by HomVEE used staff that met the qualifications required by program models, and staff participated in standard training provided to new program staff by national or university-based support offices.

Because most program models have changed over time, results of early evaluations may not represent the potential effectiveness of models as they are implemented today.

Many home visiting program models in the HomVEE review have a relatively long history, and their logic models have evolved over time. For example, NFP was first implemented in as part of a

1977 efficacy trial, PAT was launched in 1981, and HFA was launched in 1992. Most program models have changed over time as developers, researchers, and practitioners learn about what works well and what does not. Developers have typically made changes intended to strengthen programs and improve their potential for producing positive child and family outcomes as they have learned about program effects based on evaluation results and feasibility of implementation from practitioners, trainers, and technical assistance providers. Therefore, HomVEE users should be aware that evaluation results from early studies may not necessarily reflect the potential effectiveness of program models implemented according to current fidelity standards. If the program model has changed, results of prior studies could over- or underestimate potential program impacts.

HomVEE reported results of studies that compared adaptations or enhancements of program models separately from other studies of the program model.

In addition to developers' intentional changes and refinements to program models over time, practitioners may make other adaptations to program models for a variety of reasons, such as time and resource constraints, community norms, and characteristics of the target population (Dunsenbury et al., 2003). There is an inherent tension between maintaining a high degree of implementation fidelity and adapting program models in ways designed to better meet community needs (Kumpfer, Alvarado, Smith, & Bellamy, 2002; Castro, Barrera, & Holleran Steiker, 2010). Adaptations may include enrolling a target population not intended by the developer, adding or eliminating program components, using home visitors with qualifications different from those specified by the developers, translating program materials into another language, or changing program dosage.

The HomVEE review identified several studies that compared the effects of adaptations to those of the main program models.[5] For example, one study compared the effects of Healthy Steps to a Healthy Steps enhancement called PrePare, which provided additional prenatal home visits. An NFP study compared the results for families visited by nurse home visitors, the standard program model, with an adaptation that used paraprofessional home visitors.

As noted earlier, HomVEE found a lack of research on effects of home visiting models for several target populations of interest, including Native American and other racial/ethnic groups, new immigrants, and military families. It is likely that cultural or other adaptations of program models may be needed to recruit, enroll, and serve these families. Often, practitioners consult with program developers or purveyors to determine whether proposed adaptations are acceptable or violate core components of the model. Without studies that examine the effects of these adaptations, however, practitioners cannot know whether the effects will be similar to effects reported in the studies of program models (without adaptations) reviewed by HomVEE.

VI. External Validity: Generalizing Beyond the Study Populations

The HomVEE review focused on internal or causal validity and did not consider external validity in the ratings or results. Internal validity is key to determining whether a program is effective

[5] Because the focus of these studies are adaptations, rather than the actual program models, ratings and results of these studies did not factor into the assessment of effectiveness of the program models. The results of the studies, however, are reported on the HomVEE website.

but does not address whether a study's findings would hold with a different sample, treatment providers, or community context, known as external validity. For example, findings from a nationally representative sample are presumed to represent what would be observed in another nationally representative sample or in the entire nation's population if it were observed at that point in time. Assessing the extent or strength of external validity is difficult, however, because of a lack of well-established or commonly used standards.

For the HomVEE review, two key issues limit the external validity of the findings: the time frame in which some studies were conducted and the study samples.

Older research is necessary for longitudinal follow-up that examines long-term effects, but also is dated.

HomVEE examined studies conducted within a 30-year time frame, 1979 to 2009, to ensure that all trials of prevalent home visiting models were included and to include studies with long-term follow-up. Findings from older research, however, may not represent what would be observed today. The counterfactual and context in which programs and evaluations are implemented are always changing. Service delivery practices that were effective in the past may no longer be effective due to changes in cultural norms, family needs, or widespread availability of similar services in the community. Often, as conditions and the availability of services improve, effects observed in the past are attenuated. In some ways, program models may appear to be victims of their own success, as effective services or programs are adopted by other organizations.

At the same time, research conducted earlier is necessary for observing sustained findings over time. To measure long-term results, time must pass and researchers must continue to follow families and children enrolled in studies begun years earlier. For example, the Elmira, New York, NFP study, which began in the late 1970s, followed children through age 15, allowing for the detection of possible effects until adolescence. Although the study provides important information about long-term effects, the counterfactual may be very different today because of new programs and services available to children and families. This is a necessary tension when considering older research—balancing the importance of learning about long-term effects with considerations of whether the program effects represent what would occur in today's context.

Most studies in the home visiting literature are based on selected samples that limit the generalizability of the findings.

The studies reviewed for HomVEE typically had limited external validity. Many had small samples of convenience rather than samples selected to be representative of the eligible target population for the program model. In addition, some studies were conducted in a single location whose context might not translate to other types of communities or regions of the country. Although the positive results of these studies are promising, it is difficult to predict whether they would be replicated with other groups of children and families and in other communities.

Some models, such as NFP and HFA, had multiple trials and samples examining the effects of the programs. Although the individual samples were typically homogeneous, the use of multiple samples provides more confidence that the findings were not limited to a single location or sample.

From a practical standpoint, obtaining a representative sample can be extremely expensive and is often beyond the capacity of the available funding. Further, even if a sample is representative, the study may be limited in its ability to generalize to other types of service providers or to adaptations made for different kinds of families. For HomVEE users, it may be useful to consider details of specific studies, especially the context for implementation and the characteristics of the research sample, to assess the likelihood that the model is a good fit in a specific community and has the potential to produce positive outcomes for the intended population.

VII. Addressing Conflicts of Interest: Is the Research Free of Bias?

Conflicts of interest can call study results into question. Potential conflicts can occur when research is conducted or funded by those who have a vested interest in the study findings. Although these conflicts do not necessarily lead to biased results, they can cause consumers of the research to question the findings. Examples of potential conflicts include study funding provided by a curriculum publisher who stands to gain financially from positive findings or involvement of a program developer in an evaluation. Although it is common for program developers to be involved in early efficacy trials of prevention interventions, consumers of research can have more confidence in evidence of effectiveness that includes later replication studies conducted by researchers other than the program developers. The HomVEE review did not consider conflicts of interest in assigning study ratings or program model ratings. However, the review documented the funding source for each study reviewed and noted whether any of the study authors were program developers. Nearly all articles published on NFP and Family Check-Up included a program developer as an author. None of the high- or moderate-quality studies for other program models reviewed were authored by program developers.

VIII. Summary of Lessons Learned from the HomVEE Review

Based on the first year of the HomVEE review, we have gleaned the following lessons about the state of the early childhood home visiting research field:

Research Designs

- Most prioritized models had at least one RCT that met HomVEE standards.
- None of the QEDs (matched comparison, single case, or regression discontinuity designs) reviewed by HomVEE met standards for a high or moderate rating.
- HomVEE found few single case and no regression discontinuity designs in the home visiting research literature.
- Although some program models were the subject of many publications, the analyses were typically based on three or fewer analytic samples.

Assessing Effects

- Studies in the home visiting literature use a wide range of outcome measures of varying quality, making it difficult to compare findings across studies.
- Home visiting evaluations typically measure numerous outcomes, both within and across domains as well as over time, yet few correct for multiple comparisons.

- Some studies did not report standard effect sizes, or provide sufficient information for their calculation, creating challenges for making comparisons across studies and program models.
- It is not feasible to make judgments about the meaning of effect sizes across outcome domains because the policy and clinical relevance of effects across domains varies widely.
- Some program models have evidence of sustained effectiveness one or more years after the end of the program enrollment period, but most studies did not measure longer-term effects after program completion.
- The home visiting literature shows some consistency in favorable findings, within or across studies, however, in most studies, the majority of outcomes did not show statistically significant differences. Thus, favorable or unfavorable findings were typically not part of a larger pattern of statistically significant findings.

Effects on Subgroups

- More research is needed on the effect of program models for specific target populations, including racial and ethnic groups such as Native Americans and Hispanics, recent immigrants, and military families, as well as on the effects of program adaptations that may be needed to serve these families.
- Most subgroup findings in the home visiting research literature are not replicated in multiple samples and studies are likely to be underpowered to detect subgroup effects.

Interpreting Results

- Most studies do not include assessments of implementation fidelity.
- Because most program models have evolved and the context for program implementation has changed over time, early evaluations may not reflect the potential effectiveness of models as they are implemented today.
- Most studies of home visiting models are based on selected samples that limit their generalizability. A few program models have replication studies based on different samples and locations, which increases confidence in the findings.
- Studies for some program models have been conducted with substantial involvement by program developers, creating potential conflicts of interest.

IX. Suggestions for Future Research

The growing body of home visiting research literature has a number of strengths. Most program models reviewed by HomVEE had at least one RCT that met HomVEE standards. Multiple rigorous and high-quality studies exist for some programs models. Limitations and gaps in the literature continue, however, especially with regard to replication studies, outcome measurement, fidelity assessment, research on subgroups, and researcher independence. Future research can build on the foundation established by extant work and strengthen the growing evidence base for home visiting models. Based on the first year of the HomVEE review, we offer the following recommendations:

Use research designs with strong internal validity.

These designs include RCTs, RD designs, and SCDs. However, other characteristics of the designs can weaken this validity, such as reassignment (in the case of RCTs and RD designs) and an insufficient number of data points (for SCDs). HomVEE and other reviews, including the Pregnancy Prevention Research Evidence Review, the What Works Clearinghouse, SAMSHA's National Registry of Evidence-Based Programs, Campbell Collaboration, and Blueprints, offer guidelines on constructing and implementing rigorously designed studies.

The HomVEE review results suggest that the use of RCTs is more common than RD designs or SCDs. These other designs, however, may be less expensive and more feasible to implement and should be considered for future research. In an SCD, for example, each case will receive the treatment, and thus treatment is not withheld from any sample members during the evaluation, as must be done for RCTs.

Other matched comparison designs cannot definitively rule out the possibility of selection bias, but they can provide valuable information on program effectiveness if well implemented. In addition, matched comparisons may be one of the easier designs to implement because selection into treatment and comparison conditions can be purposeful. To represent a reasonable counterfactual, however, it is imperative that the treatment and comparison groups are similar on observed characteristics at baseline.

Select study samples with external validity in mind.

Researchers and practitioners generally are interested in the model's effectiveness beyond any given study sample. An externally valid sample is representative of a population, such as all those eligible for services in a state, a region, or a neighborhood, which requires taking a random sample so that every member of a population has a chance of being included in the study. External validity also may apply to the types of providers delivering the services, community context, or other factors.

When designing a study, researchers may want to think carefully about the population of interest and try to construct a study that represents that population. To conduct a random sample requires a sampling frame, which will document every member of that population; this approach may not be feasible. However, other options should be considered—for example, enumerating and sampling a smaller population of interest, such as a neighborhood. If true representativeness is not feasible, researchers can still ensure variation in the sample of providers and participants and inclusion of target populations and characteristics of interest, such as different types of service providers, race/ethnicity and languages spoken by families, regions of the country, and community characteristics.

Build plans for assessing potential effects on subgroups into the research design.

Program effects on a subgroup should be considered before a sample is drawn. For example, are there reasons to believe a model may be more (or less) effective for a certain population, such as teen parents, particular racial/ethnic groups, or military families? If so, then a sample should be drawn that has adequate power to detect differences for that subgroup. This often requires oversampling the group of interest, to ensure there are sufficient sample members with the characteristic(s) of interest. By planning for the eventual analysis of a subgroup, a study can prevent the problem of not detecting a finding because the sample is too small. Further, selecting a subgroup beforehand can help avoid the issues associated with data mining and multiple comparisons described earlier. A subgroup that is defined *a priori* is preferable to one that is selected because a statistically significant finding for a particular group was uncovered through multiple tests.

Determine the appropriate sample size to detect statistically significant findings of interest.

Underpowered studies miss associations that should be statistically significant. More formally, power is the probability that a statistical test will correctly reject the null hypothesis of no effect when it is false (Shadish et al., 2002). Thus an underpowered study will not detect statistically significant associations even when the relationship exists. There are a number of ways to increase the statistical power of a study, such as increasing the sample size and measuring covariates that will increase the precision of the analysis. Determining whether a study is adequately powered requires a number of considerations, such as the expected effect size of the program model, but many computer programs can estimate the power of a sample using these assumptions.

Select a focused set of outcome measures that are closely aligned to the program model's targets of change, have strong validity and reliability, are appropriate for the study population, and allow for cross-study comparisons.

As noted earlier, home visiting studies typically measure outcomes in a wide range of domains and use multiple measures within domains. Using a more focused set of measures with strong validity and reliability can increase confidence in measurement accuracy and make patterns of findings more apparent. Studies can be strengthened by selecting measures that are closely aligned to the program model's theory of change and hypothesized outcomes. High-quality measures, such as those classified as primary measures in the HomVEE review, are preferred if available and feasible. In general, direct assessment or observation is preferable to self reports based on nonstandardized measures. In some cases, however, appropriate primary measures may not be available in a family's home language, or may not be feasible due to cost. Moreover, administrative records, such as reports of child maltreatment, may not be reliable sources of information in some domains. In addition, measures should be culturally appropriate for the study sample and available in families' home languages. Researchers should also select measures with comparability in mind. Consistent outcome measurement across studies of the same program model, as well as across models to the extent feasible, can facilitate detection of patterns of impacts as well cross-model comparisons.

Measure longer-term effects of the program model.

If a home visiting model intends to have sustained impacts that last after the program has ended, these effects should be measured. Researchers and developers will need to carefully consider what length of follow-up is reasonable. The program model's theory of change and expectations about longer-term effects can be used as a guide for making this decision. Researchers must also balance the need for capturing longer-term outcomes with the resources needed to track a sample over time and measure follow-up outcomes at multiple points in time. Typically, attrition increases over time, which can compromise designs with strong internal validity. However, longer-term follow-up is the only way to determine whether a model's effects are sustained.

Assess implementation fidelity as part of outcome evaluations.

Studies that do not measure implementation fidelity run the risk of drawing incorrect conclusions about the effectiveness of a program model if it was not implemented according to the developer's specifications. Especially in early efficacy trials, assessments of fidelity are also useful for determining the feasibility of implementing the program as intended. In later effectiveness studies, measuring implementation fidelity can provide information about levels of fidelity likely to be obtained in real-world settings in which developers are not involved in monitoring implementation.

Take steps to reduce the risk of finding statistically significant findings by chance when conducting multiple comparisons.

Most home visiting studies measure outcomes in multiple domains, so steps should then be taken to reduce the likelihood of finding statistically significant findings by chance. Corrections can be made, such as Bonferroni, which adjust the alpha levels to account for multiple tests. Critics, however, contend that these corrections are too severe and lead to an increase in Type II error (the probability of not detecting a statistically significant finding).

Another possibility for addressing this issue is selecting key or confirmatory variables of interest that are the focus of the program model. Thus if the model targets the reduction of child maltreatment, then this could be considered a primary outcome, whereas other outcomes, such as family self-sufficiency, may be less important. Multiple comparison corrections are then only applied to key outcomes (for example, multiple indicators of child maltreatment). Secondary or exploratory outcomes do not require corrections, but it is understood that the conclusions regarding statistical significance are more tentative and do not represent the primary test of whether the model was successful in achieving its key objectives. A statistically significant exploratory outcome in one study may become the primary outcome of a replication study. With this approach, the outcomes should be selected *a priori*, before any data analysis. This precludes the possibility of data mining, that is, hunting for significant associations and then declaring those the confirmatory outcomes.

Medical research offers a model for tracking hypotheses and the selection of confirmatory and exploratory outcomes. RCTs are registered before their start (www.clinicaltrial.gov), outcomes are described and categorized as primary or secondary, and other details are provided. Furthermore, some medical journals will only publish results from registered trials. This system increases confidence in published results by ensuring that researchers neither drop outcomes that did not show the desired effects nor publish only partial results.

Report effect sizes.

Because effect sizes show the size of the impact relative to the standard deviation of the measure and are independent of the units in which the outcome in measured, they facilitate comparisons of results across outcomes and studies. Interpreting the meaning of effect sizes—what range of effect sizes have clinical or policy relevance—is challenging. Each outcome domain should be considered on its own rather than in comparison to others because the meaning of relative effect sizes will vary across outcome domains. Providing context for effect sizes can aid in their interpretation, such as considering the size of normative changes in absence of an intervention (Hill et al. 2008). For example, how does the effect of an intervention compare to children's typical cognitive development over the same period of time as the study's follow ups (Hill et al. 2008)? Effect sizes also can be compared to those of similar interventions, but caution must be used if the reported outcomes differ.

Conduct studies with multiple study samples that seek to replicate the findings of initial efficacy trials.

The replication of favorable findings increases confidence in the effectiveness of a program model. Replication shows that the findings were not an isolated result limited to a single sample or due to heavy involvement by the program developer, as is typical in initial efficacy studies. Replication may not have the same appeal as an initial trial because it is by definition confirming past results. Accordingly, funding for replications may be more difficult to secure. Nevertheless, replication is key for confirming findings from early studies, particularly those that were unexpected, results based on subgroups that were not established ahead of time, or findings that did not have the appropriate corrections for multiple comparisons. Ideally, replication studies should be conducted by a research team that is independent from the original. Replication studies should be based on a different analytic sample than was the original but should use the same outcome measures to the extent feasible to allow for comparisons across studies. Follow-up studies of the same sample do not constitute replication.

Continue to test the effectiveness of the program model periodically, as earlier results may be less applicable to today's families and context.

Both the program model and the counterfactual are likely to evolve and change over time. Models may modify components based on lessons learned from past evaluations or feedback from practitioners. Further, as successful approaches to service delivery are disseminated and replicated, the counterfactual—what would happen in absence of the program—changes. Therefore, ideally, research on a model should continue, not just to replicate past results but also to ensure that the results reflect the current environment and needs of children and families.

REFERENCES

Boller, K., Strong, D., & Daro, D. (2010). Home visiting: Looking back and moving forward. *Zero to Three, 30*(6), 4–9.

Brown, C. H. (2009). “Foundational Issues in Examining ‘Subgroup’ Effects in Experiments.” Paper presented at the Interagency Federal Methodological Meeting: Subgroup Analysis in Prevention and Intervention Research, Bethesda, Maryland, September 14-15.

Caldwell, B., & Bradley, R. (1984). *Home observation for measurement of the environment (HOME)* (Rev. ed.). Little Rock: University of Arkansas.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40). Retrieved November 30, 2010, from http://www.implementationscience.com/content/2/1/40

Castro, F. G., Barrera, M., & Holleran Steiker, L. K. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology, 6*, 213–239.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23–45.

Daro, D. (2006). *Home visiting: Assessing progress, managing expectations.* Chicago, IL: Chapin Hall at the University of Chicago.

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3,* 269–276.

Dunsenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256.

Gomby, D. S. (2005). *Home visitation in 2005: Outcomes for children and parents* (Invest in Kids Working Paper No. 7). Washington, DC: Committee on Economic Development.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives, 2*(3), 167–177.

Hill, C. J., Bloom, H. S., Rebeck Black, A., Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.

Klerman, J. A. (2009). “Subgroup Analysis: A View from the Trenches.” Paper presented at the Interagency Federal Methodological Meeting: Subgroup Analysis in Prevention and Intervention Research, Bethesda, Maryland, September 14-15.

Knoche, L. L., Sheridan, S. M., Edwards, C. P., & Osborn, A. Q. (2010). Implementation of a relationship-based school readiness intervention: A multidimensional approach to fidelity assessment for early childhood. *Early Childhood Research Quarterly, 25*, 299–313.

Kumpfer, K. L., Alvarado, R., Smith, P., & Bellamy, N. (2002). Cultural sensitivity and adaptation in family-based prevention interventions. *Prevention Science, 3*(3), 241–246.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Nurse Family Partnership. (2010). Home page. Retrieved November 3, 2010, from http://www.nursefamilypartnership.org/

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33–84.

Paulsell, D., Avellar, S., Sama Martin, E., & Del Grosso, T. (2010). *Home visiting evidence of effectiveness: Executive summary.* Princeton, NJ: Mathematica Policy Research.

Parents as Teachers. (n.d.). Vision | mission history. Retrieved November 3, 2010, from http://www.parentsasteachers.org/about/what-we-do/visionmission-history

Pew Center on the States. (2010). *Pew inventory of state home visiting programs.* Retrieved November 30, 2010, from http://www.pewcenteronthestates.org/initiatives_detail.aspx? initiativeID=61051

Scanlon, J.W., Horst. P, Nay, J. N., Schmidt, R. E., & Waller, J. D. (1979). Evaluability assessment: Avoiding type III or IV errors. In G. R. Gilbert, & Y. P. J. Conklin (Eds.), *Evaluation management: A source book of readings.* Washington, DC: Office of Personnel Management, Federal Executive Institute.

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review, 33,* 539–567.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin.

Society for Prevention Research. (n.d.). *Standards of evidence: Criteria for efficacy, effectiveness, and dissemination.* Retrieved December 1, 2010, from http://www.preventionresearch.org/StandardsofEvidencebook.pdf

Stoltzfus, E., & Lynch, L. (2009). *Home visitation for families with young children.* Washington, DC: Congressional Research Service.

Wasik, B. H., & Bryant, D. M. (2001). *Home visiting procedures for helping families* (2nd ed.). Thousand Oaks, CA: Sage.

MATHEMATICA
Policy Research, Inc.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC