Home Visiting Evidence of Effectiveness (HomVEE) Systematic Review Handbook of Procedures and Evidence Standards: Version 1

December 2020

OPRE Report 2020-169

Home Visiting Evidence of Effectiveness (HomVEE) Systematic Review Handbook of Procedures and Evidence Standards: Version 1

OPRE Report 2020-169

Submitted to:

Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
330 C St SW, 4th Floor, Washington, DC 20201

Contract Number: GS-10F-0050L/HHSP233201500115G

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221

Contents

Exhibits

Introduction

The Home Visiting Evidence of Effectiveness (HomVEE) systematic review is funded by the Department of Health and Human Services (HHS) and is designed to be a thorough and transparent review of the research literature on early childhood home visiting. HomVEE, which began in 2009, is an assessment of evidence of effectiveness, as defined by HHS, for home visiting models that serve families with pregnant women and children from birth to kindergarten entry (up through age 5).

The original procedures and standards for the review were developed with guidance from a Department of Health and Human Services (HHS) interagency working group composed of representatives from:

- The Office of Planning, Research, and Evaluation (OPRE), Administration for Children and Families (ACF)
- The Children's Bureau, ACF
- The Centers for Disease Control and Prevention (CDC)
- The Health Resources and Services Administration (HRSA)
- The Office of the Assistant Secretary for Planning and Evaluation (ASPE)

HomVEE provides information about which home visiting models have evidence of effectiveness as defined by HHS as well as 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.

This document describes the original process and standards HomVEE used to review research from impact studies. In 2020, HomVEE released a Handbook of Procedures and Standards: Version 2. Those newer procedures and standards, which can be found on the HomVEE website alongside this original version (https://homvee.acf.hhs.gov/publications/methods-standards), are in effect beginning with the 2021 annual review and until any subsequent version of the procedures and standards.

Overview

To conduct a thorough and transparent review of the home visiting research literature, HomVEE follows seven main steps:

1. Conduct a broad literature search.
2. Screen studies for relevance.
3. Prioritize models for the review.
4. Rate the quality of impact studies with eligible designs.
5. Assess the evidence of effectiveness for each model.
6. Review implementation information for each model.
7. Address potential conflicts of interest.

To have a complete understanding of possible program effects, the review must include all relevant research to date on models. Thus, reviews of new models and updates of existing models systematically include all of the aforementioned steps.

As of 2019, HomVEE divides reviews into two tracks:

- Track 1 is for models that HomVEE has not previously found to be evidence based (that is, models that either have never been reviewed by HomVEE or were reviewed but did not meet the criteria for evidence of effectiveness). HomVEE releases results for models in Track 1 in September of each year.
- Track 2 updates the review of literature on models that HomVEE has previously found to be evidence based. HomVEE releases results for models in Track 2 in December of each year.

HHS created these two tracks to facilitate review of a greater volume of models that are not evidence based while still keeping reviews of evidence-based models up to date.

Literature Search

Each October, the HomVEE team conducts a broad search for literature on home visiting models serving pregnant women or families with children from birth to kindergarten entry (that is, up through age 5). The team limits the search to research on models that used home visiting as the primary service delivery strategy and offered home visits to most or all participants. The search is also limited to research on home visiting models that aim to improve outcomes in at least one of eight domains: (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. HomVEE's literature search includes two main activities:

1. **Database searches.** The HomVEE team searches on relevant key words in a range of research databases. Keywords include terms related to the service delivery approach, target population, and outcome domains of interest. The HomVEE team also performs focused searching, by model name, for models with the highest prioritization scores in each year. The initial search was limited to studies published since 1989; a more focused search on prioritized models included studies published since 1979 (see the section on Prioritizing Models for Review, later in this document). This search is updated annually to identify new literature released between October and the end of September. (Before the 2019 review, HomVEE's search covered January through December of the previous year.) HomVEE also considers submissions to the annual call for studies, which is released in November and open through early January of the review year.

Exhibit 1. Keywords used in the Home Visiting Evidence of Effectiveness Literature Review

Category	ID	Search term
Search restrictions	--	Studies published in English only Studies published after 1989
Activity	S1	(Home AND visit*) or "family development" or (case AND manage*) or ((coordination OR referral*) AND (home AND visit*))
Target group	S2	Prenatal or perinatal or pregn* or "early childhood" or preschool or "pre-school" or infan* or newborn* or toddler* or parent* or "low-income" or "low income" or poor or poverty or "young child*"

Category	ID	Search term
Outcomes	S3	(child* and (abuse or neglect or maltreatment or health or injury or violence or attachment or immuniz* or "emergency department")) or "infant mortality" or ((juvenile or adolescent) AND delinquen*) or (child and (cognit* or language or "social-emotional" or "socioemotional" or "socio-emotional" or socioemotional or physical or health) and development) or "school readiness" or "school achievement" or "child development" or "developmental delay" or (child AND behavior*) or (child AND disab*) or ((preterm or "pre-term" or premature) AND birth) or "low birth weight" or ((parent* or family or matern* or mother* or father* or patern*) and (employment or career or stress or depress* or efficacy or "mental health" or health)) or ((subsequent or teen) AND (birth or pregnan*))or "home environment" or (parent* AND (skill* or abilit*)) or (reduc* AND (crime or "domestic violence" or "family violence" or "intimate partner violence")) or ((community AND coordinat* OR co-ordinat*) or referral*) or "self sufficiency" or "self-sufficiency" or (smoking or tobacco) or ("armed forces" or military)
Document type	S4	(Study or evaluat* or research) and (effective* or efficac* or impact* or outcome* or implement* or cost or replic*)
Combine terms	S7	S1 AND S2 AND S3 AND S4

Databases searched

- Academic Search Premier
- Campbell Collaboration
- Child Care & Early Education Research Connections
- CINAHL with Full Text
- Cochrane Central Register of Controlled Trials
- Cochrane Methodology Register
- Database of Abstracts of Reviews of Effects
- E-Journals
- EconLit
- Education Research Complete
- ERIC
- MedLine
- New York Academy of Medicine's Grey Literature Report
- ProQuest Dissertations (prior to 2012, used Dissertation Abstracts)
- PsycINFO
- Scopus (prior to 2013, instead searched Social Sciences Citation Index, Social Work Abstracts, and Sociological Abstracts)
- SocINDEX with Full Text
- WorldCat

2. **Call for studies.** Each fall, HomVEE issues a call for studies and sends it to relevant electronic mailing lists for dissemination. The call for studies closes in January and may include unpublished studies or studies published through December of the previous year. Submissions to the call for studies are added to HomVEE's database of studies, along with studies obtained through the database searches. Each year, HomVEE screens all studies for relevance and prioritizes models for review; effectiveness and implementation studies on prioritized models are reviewed throughout the rest of the calendar year. HomVEE retains all submissions that are eligible for review. However, because of the volume of research, HomVEE cannot fully review all submitted studies. Studies that are not about one of the prioritized models are included for consideration in subsequent review cycles.

In addition to these two activities, in the first year of the review, HomVEE also included the following:

3. **Review of existing literature reviews and meta-analyses.** In the first year, the HomVEE team checked initial search results against the bibliographies of recent literature reviews and meta-analyses of home visiting models and added relevant missing citations to the search results. This check ensured that our search terms identified relevant studies; after confirming the validity of the search terms, we did not repeat the process in subsequent years.
4. **Website searches.** The HomVEE team used a custom Google search engine to search more than 50 relevant government, university, research, and nonprofit websites for unpublished reports and papers. However, results of this search largely overlapped with the results of the first two activities, so the team discontinued this activity in subsequent years.

Screening Studies

Each year, the HomVEE review team screens all new citations identified through the literature search for relevance. The team screens out studies for the following reasons:

- Home visiting was not the primary service delivery strategy.
- The study did not use an eligible design. (Eligible designs are randomized controlled trials, quasi-experimental designs, and implementation studies.)
- The study of the program did not report results for an eligible target population: pregnant women or families with children from birth to kindergarten entry (that is, up through age 5) served in a developed-world context.
- The study did not examine any outcomes in the eight eligible outcome domains (child development and school readiness; child health; family economic self-sufficiency; linkages and referrals; maternal health; positive parenting practices; reductions in child maltreatment; and reductions in juvenile delinquency, family violence, and crime).
- The study did not examine a named home visiting model.
- The study was not published in English.
- The study was published before 1989.[1]

[1] For models prioritized in 2018 and earlier, HomVEE also did a focused search reaching back to 1979. Because so few studies published before 1979 related to models prioritized in recent years, HomVEE limited the focused search to studies reaching back to 1989 or later starting with the 2019 review. For the 2019 review, HomVEE searched literature published through September 2018. We also considered submissions of unpublished studies or studies published through December 2018 to the call for studies that ended in early January 2019.

Prioritizing Models for Review

Each year, HomVEE releases new review results for models. This includes reviews of studies on models that have not previously been reviewed, updates to previously reviewed models, or both. Decisions on the number of models to review depend on (1) the number of studies that are identified for review about each model and (2) the available project resources. The process by which HomVEE selects models for review is called the **prioritization process**.

Prioritization process

HomVEE selects models for the annual review by calculating a prioritization score for each model and then reviewing models with the highest scores. The prioritization score is based on points assigned at the study and model levels. HomVEE divides reviews into two tracks. Track 1 is for models that are not evidence based (that is, models that either HomVEE has never reviewed or that HomVEE has reviewed but did not meet the criteria for evidence of effectiveness). Track 2 updates the review of literature on evidence-based models. HomVEE prioritizes models separately in each track, but the process is largely similar for both. The prioritization process has six steps:

1. Identify studies eligible for review, which includes the HomVEE literature search and screening studies activities.
2. Apply study-level criteria
3. Apply model-level criteria
4. Calculate prioritization scores
5. Adjust prioritization scores
6. Prioritize models

Below, we describe Steps 2 through 5 in the prioritization process and how it differs for models in Tracks 1 and 2.

Step 2. Apply study-level criteria. HomVEE reviews the titles and abstracts of impact studies identified for each model and assigns points based on HomVEE's prioritization criteria. This process is identical for both tracks. Models can earn up to 5.75 points for each eligible impact study (Table 1). HomVEE assesses each study (manuscript) separately and then sums the points for all studies about a model. Therefore, models with more eligible studies tend to receive more study-level points. Whether a model is already evidence based determines which studies are included in that model's study-level point total:

- If a model is not evidence based (Track 1), the total includes study-level points for studies that HomVEE reviewed in previous years and assigned a high or moderate rating as well as studies that HomVEE has not previously reviewed.
- If a model is already evidence based (Track 2), the total includes points only for studies that HomVEE has not reviewed yet.

Exhibit 2. HomVEE study-level prioritization criteria and associated points

Criterion	Points	Notes
Number and design of impact studies	2 to 3 per study	3 points for each randomized controlled trial, single-case design, or regression discontinuity design 2 points for each matched-comparison group design[a]
Sample size	1 per study	Study sample contains 250 or more pregnant women and/or families
Outcomes of interest	1 per study	Study examines outcomes in one or more of the following domains: family economic self-sufficiency; linkages and referrals; reductions in child maltreatment; and reductions in juvenile delinquency, family violence, or crime;[a]
Study sample	0.5 per study	Study sample lives in the United States or is an indigenous population in or outside of the United States
Priority population	0.25 per study	The entire sample belongs to one or more priority populations named in the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) authorizing statute[b]

Note: HomVEE applies these points at the study level based on information that study authors provide in the title and abstract. HomVEE assesses each study separately and then sums the points for all studies to create a study-level total for the model.

[a] More information about these outcomes is available at https://homvee.acf.hhs.gov/outcomes.

[b] According to 42 U.S.C. § 711 (d)(4), priority populations are as follows:

- Low-income families
- Families with pregnant women who have not reached age 21
- Families that have a history of child abuse or neglect or have had interactions with child welfare services
- Families that have a history of substance abuse or need substance abuse treatment
- Families that have users of tobacco products in the home
- Families that are or have children with low student achievement
- Families with children with developmental delays or disabilities
- Families that include individuals who are serving or formerly served in the Armed Forces, including such families that have members of the Armed Forces who have had multiple deployments outside of the United States

Step 3. Apply model-level criteria. HomVEE assigns model-level points based on information from study titles and abstracts, model websites, and previous HomVEE reviews. HomVEE may contact study authors or model developers to confirm publicly available information. This process is identical for Tracks 1 and 2 (models that are and are not evidence based). Models can earn up to 4 points in this step:

- 1 point if the model is associated with a national organization or institution of higher education (organizations can be in or outside the United States)
- 1 point if the model is currently serving or available to serve families
- 1 point if the model has been implemented for at least three years (even if it is not currently active)
- 1 point if support is available to implement the model in the United States

These model-level factors are specific MIECHV Program-relevant criteria that are intended to more closely align HomVEE with the MIECHV Program.

Step 4. Calculate prioritization scores. After assigning study- and model-level points, HomVEE sums all points to calculate a model's point total. For models that are not yet evidence based (Track 1), the total is the final model prioritization score. For models that are evidence based (Track 2), there is one additional step: HomVEE assigns a weight to the Track 2 model score based on the number of years since HomVEE last reviewed the model and released a report. The weight is calculated using the following formula:

Weight = $[1 + 0.1 * (\text{current year} - \text{release date of prior report})]^2$

For example, a model considered for review in 2019 for which its most recent HomVEE report was released in 2015 would receive a weight of $[1 + 0.1 * (2019 - 2015)]2 = 1.96$.

After calculating weights, a model's final prioritization score is then calculated as:

Prioritization score = Model point total * Weight

The weights permit Track 2 (evidence-based) models with new research to be updated periodically. Models that were reviewed longer ago have a higher weight than models reviewed more recently. This increases the relative likelihood that a model that has not been reviewed recently will be prioritized for review.

Step 5. Adjust prioritization scores. After calculating prioritization scores, HomVEE sorts models from highest to lowest score separately within each track. The team then conducts a second, focused database search on model names to identify additional studies about top-scoring models in each track. The model's prioritization score is adjusted to add the study-level points for the newly identified studies. Additionally, HomVEE examines the full texts of all screened-in studies about top-scoring models and then adjusts the study-level point totals (and the model's corresponding prioritization score) based on information available from the full text.

Step 6. Prioritize models. HomVEE re-sorts models from highest to lowest using the adjusted prioritization scores and identifies models with the highest scores. After prioritizing a model, HomVEE reviews all new impact studies about that model, with two exceptions:

- In years when resources are limited, HomVEE will not review research conducted outside the United States if it is about a model that is already evidence based. (Research with indigenous communities outside of the United States will still be reviewed.) If studies conducted outside of the United States are not reviewed, the model report on the HomVEE website will clearly indicate which research was and was not included in the updated report.
- Regardless of whether the model has previously been found to be evidence based, HomVEE will not update results for previously reviewed models every year. Models are only considered for updates every two years at the earliest. For example, if review results for a model were updated in 2019, that model would not be considered for additional updating until 2021 or later.

HomVEE's prioritization process reflects HomVEE's emphasis on identifying new evidence-based home visiting models while continuing to update reports on models that are already evidence based. The annual prioritization effort may yield more models in the highest point category than can be reviewed that year. The number of models reviewed each year depends on the available project resources and the number of studies identified to review for each model. Regardless of whether a model is reviewed in a given year, all models will be included in the prioritization process in subsequent years. The MIECHV Program may

coordinate with HomVEE to prioritize review of promising approaches implemented and evaluated under a MIECHV Program grant.[2]

Producing Study Ratings

A. Reviewing eligible studies

Trained reviewers evaluate randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) identified for each prioritized model, assessing the research design and methodology of each study using a standard review protocol. To ensure the accuracy of reviews, each study is reviewed by two members of the review team. The first reviewer evaluates the study, assigns a study rating (see below), and records the review results. A second reviewer examines the study and the results of the first review. If the second reviewer disagrees with any of the first reviewer's decisions, the two reviewers discuss these differences to reach a consensus rating. The principal investigator or a senior reviewer confirms all consensus rating decisions.

Some studies are missing information related to the study rating, such as information on attrition or baseline equivalence of the treatment and comparison groups. In these cases, the HomVEE review team sends queries to authors to request the missing information. Authors are given one week to respond, but the HomVEE team includes information sent after the deadline whenever feasible. If the authors do not respond or provide the necessary information, a HomVEE reviewer assigns rating based on the available information.

If a study has a design for which HomVEE does not have existing standards, it will not be reviewed, nor will it contribute to the model's evidence base.

B. Study ratings

We review and rate studies identified in the literature search that examine the impact of a home visiting model using quantitative data and statistical analyses. The study-level ratings—(1) high, (2) moderate, and (3) low—provide a measure of how well the study design could provide unbiased estimates of model impacts. In brief, the high rating is reserved for random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment, and single-case and regression discontinuity designs that meet the What Works Clearinghouse (WWC) design standards (Table 2). The WWC, established by the Institute of Education Sciences in the U.S. Department of Education, reviews education research. The moderate rating applies to random assignment studies that, due to flaws in the study design or analysis (for example, high sample attrition), do not meet all the criteria for the high rating; matched comparison group designs; and single case and regression discontinuity designs that meet WWC design standards with reservations. Studies that do not meet all of the criteria for either the high or moderate ratings are assigned the low rating.

[2] Under federal law, a home visiting service delivery model that qualifies as a promising approach conforms to a "promising and new approach" to achieving specified benchmark areas and participant outcomes; has been developed or identified by a national organization or institution of higher education; and will be evaluated through a well-designed and rigorous process (see Social Security Act, Title V, § 511 (d): https://www.ssa.gov/OP_Home/ssact/title05/0511.htm).

Exhibit 3. Summary of study rating criteria for the HomVEE review

HomVEE study rating	Randomized controlled trials	Quasi-Experimental Designs		
		Matched comparison group	Single-case design[b]	Regression discontinuity[b]
High	• Random assignment • Meets WWC standards for acceptable rates of overall and differential attrition[a] • No reassignment; analysis must be based on original assignment to study arms • No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods • Baseline equivalence established on tested outcomes and demographic characteristics **OR** controls for these measures	Not applicable	• Timing of intervention is systematically manipulated • Outcomes meet WWC standards for inter-assessor agreement • At least three attempts to demonstrate an effect • At least five data points in relevant phases	• Integrity of forcing variable is maintained • Meets WWC standards for low overall and differential attrition • The relationship between the outcome and the forcing variable is continuous • Meets WWC standards for functional form and bandwidth
Moderate	• Reassignment OR unacceptable rates of overall or differential attrition[a] • Baseline equivalence established on tested outcomes and demographic characteristics AND controls for baseline measures of tested outcomes, if applicable[c] • No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods	• Baseline equivalence established on tested outcomes and demographic characteristics AND controls for baseline measures of tested outcomes, if applicable[c] • No confounding factors; must have at least two participants in each study arm and no systematic differences in data collection methods	• Timing of intervention is systematically manipulated • Outcomes meet WWC standards for inter-assessor agreement • At least three attempts to demonstrate an effect • At least three data points in relevant phases	• Integrity of forcing variable is maintained • Meets WWC standards for low attrition • Meets WWC standards for functional form and bandwidth
Low	Studies that do not meet the requirements for a high or moderate rating			

Note: "Or" implies that one of the criteria must be present to result in the specified rating.

[a] For ease of presentation, some of the criteria are described very broadly. Additional details are available for single-case design standards in Appendix F of the WWC version 2.1 standards (https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v2_1_standards_handbook.pdf), and in a specific document about regression discontinuity designs (http://ies.ed.gov/ncee/wwc/Document/258).

[b] The What Works Clearinghouse (WWC), established by the Institute of Education Sciences in the U.S. Department of Education, reviews education research (https://ies.ed.gov/ncee/wwc/). The WWC standard for attrition is transparent and statistically based, taking into account both overall attrition (the percentage of study participants lost in the total study sample) and differential attrition (the differences in attrition rates between treatment and control groups).

[c] The variables that must be used to establish equivalence depend on whether (1) it is possible to collect the measure at baseline vs. (2) it is difficult or impossible to collect the measure at baseline. See section below on baseline equivalence for more details.

1. Study design

In this section, we provide more details about the study rating criteria for randomized controlled trials and matched comparison group quasi-experimental designs (QEDs) in the following categories: (1) study design, (2) attrition, (3) baseline equivalence, (4) reassignment, and (5) confounding factors. (For details about design standards for single-case designs, please refer to the WWC Single-Case Design Standards (Appendix F); more information about study design standards for regression discontinuity designs can be found in the WWC Regression Discontinuity Design Standards (Appendix E).) Authors may also wish to consult the HomVEE reporting guide for study authors, or the flowcharts that illustrate standards for commonly reviewed study designs.[3] These HomVEE-specific tools are available at https://homvee.acf.hhs.gov/publications/methods-standards.

Studies that use random assignment create two or more groups that are, on average, similar to each other at the start of the study. These studies provide strong evidence that differences in the outcomes between the treatment and control groups at the end of the study can be attributed to the intervention rather than to pre-existing differences between the groups (Shadish et al. 2002). (Designs based on functionally random assignment, or assignment to groups in ways that ensure any differences between and within groups are not systematic at the outset of the experiment—such as alternation based on last name, date of birth, or certain digits of an identification number—are also eligible for the high rating.) Therefore, studies that randomly assign subjects can receive a high rating.

Matched comparison designs with an external comparison group can achieve at best a moderate rating. In such studies, subjects are sorted into the study arms through a process other than random assignment; therefore, even if the treatment and comparison groups are well matched based on observed characteristics, they may still differ on unmeasured characteristics. It is, therefore, impossible to rule out the possibility that the findings are attributable to unmeasured group differences. The moderate rating is also possible for random assignment designs that do not meet other criteria for the high rating (that is, attrition or reassignment), as explained in more detail below.

Designs without a comparison group (for example, pre-post designs) offer no way to assess what the sample's outcomes would have been in the absence of the intervention. These study designs cannot rule out that that changes were caused by, for example, history (an event besides the treatment that could have produced the observed outcome) or maturation (participants' natural changes over time could have produced the outcome) (Shadish et al. 2002). Therefore, these studies with these designs cannot meet the criteria for either the high or moderate ratings.

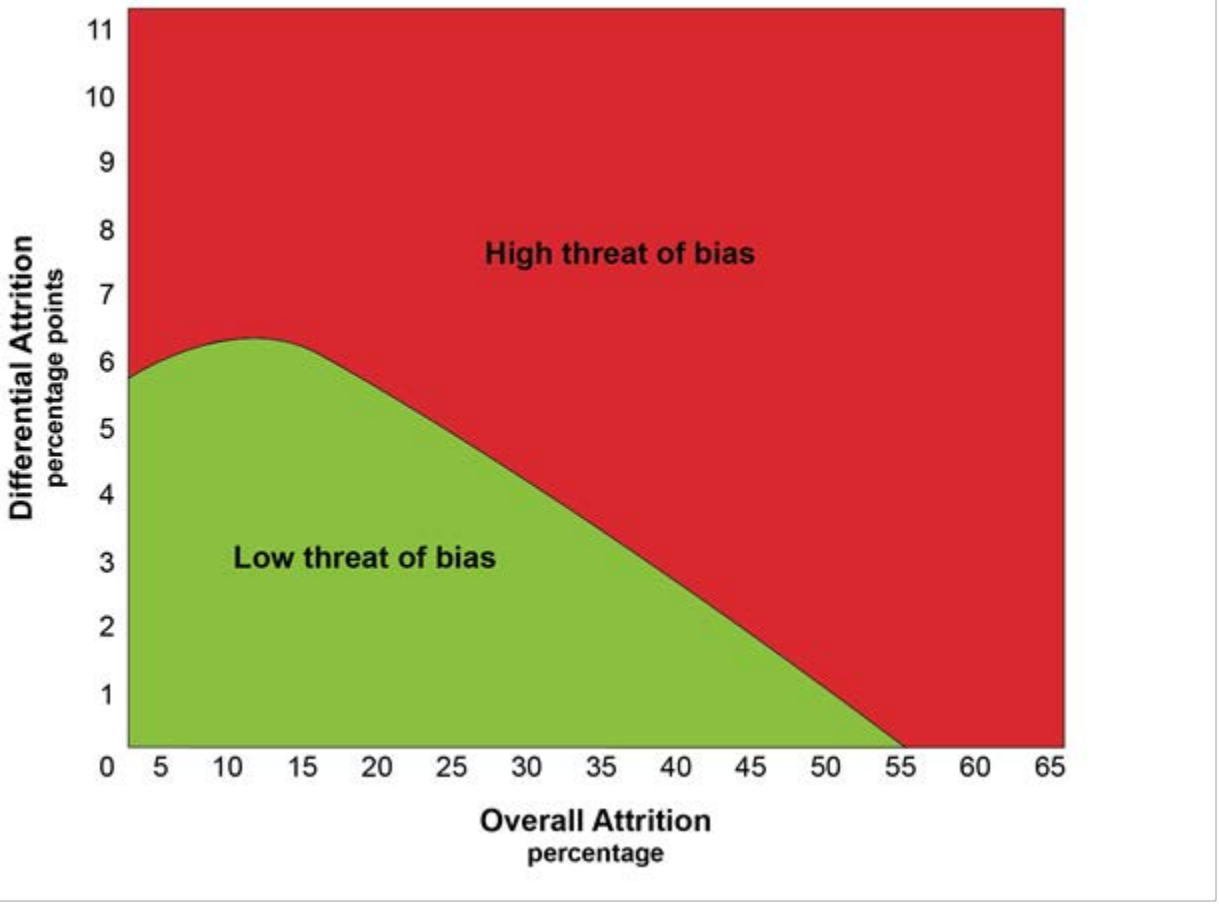
2. Attrition

In random assignment studies, a loss of study participants can bias the impact estimates by creating differences in the characteristics of the treatment and control groups. If people in the treatment and comparison groups who remain in the study were initially different from one another, posttest outcomes could differ even in the absence of treatment (Shadish et al. 2002). The HomVEE review uses the WWC standard for attrition.[4] The WWC standard for attrition is transparent and statistically based, taking into account both overall attrition (the percentage of study participants lost in the total study sample) and differential attrition (the differences in attrition rates between treatment and control groups). It recognizes

[3] Prior to 2020, HomVEE referred to non-experimental comparison group designs as matched comparison group designs. Terminology updates are reflected in the HomVEE reporting guide.
[4] https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v2_1_standards_handbook.pdf.

an important trade-off between overall and differential attrition—namely, that for an expected level of bias, studies with a relatively low level of overall attrition can tolerate a relatively high level of differential attrition, whereas studies with a relatively high level of overall attrition require a lower level of differential attrition.. The WWC attrition standard classifies studies as having either high or low attrition based on a combination of overall and differential attrition (see Exhibit 4).

Exhibit 4. Cutoffs for attrition standards



Note: The red area indicates combinations of overall and differential attrition that produce a rating of high attrition. The green area indicates combinations that produce a rating of low attrition.

Random assignment studies that meet the standard for low attrition are considered for the high study rating. Random assignment studies with high attrition are considered for the moderate study rating and must meet the other criteria for this rating.

For clustered random assignment designs in which a cluster or group, such as a neighborhood, is assigned to treatment or comparison conditions, attrition is assessed at two levels. Attrition must be low at both the cluster and individual levels to receive a high rating. If attrition is high at either or both the cluster and individual levels then baseline equivalence must be established, and the highest possible rating is a moderate. These are the same attrition standards as used by the WWC.

Exhibit 5. Attrition standards for cluster randomized trials

Level of sample attrition		Highest possible study rating
Cluster level	Individual level	HomVEE
High	Low	Moderate, with evidence of baseline equivalence **and** controls for baseline measures of outcomes
High	High	Moderate, with evidence of baseline equivalence **and** controls for baseline measures of outcomes
Low	Low	High, with evidence of baseline equivalence **or** controls for baseline measures of outcomes **as well as** race/ethnicity, and SES
Low	High	Moderate, with evidence of baseline equivalence **and** controls for baseline measures of outcomes

The attrition standards do not apply to quasi-experimental comparison group studies. These studies are evaluated on the basis of the final analysis sample, from which there is no attrition (by definition).

3. Baseline equivalence

A comparison group is intended to represent what would have happened to the treatment group in absence of the treatment. To provide the strongest evidence of this counterfactual, the treatment and comparison groups should be as similar as possible at the study's onset (that is, baseline). When the treatment and comparison groups are dissimilar, the results cannot support causal conclusions about the differential effect of the treatment (Rubin 1997).

To consider studies using matched comparison group designs and RCTs with high attrition for a moderate rating, the HomVEE review requires that:

1. The study establishes baseline equivalence on: (1) race and ethnicity, (2) socioeconomic status (SES) and (3) baseline measures of outcomes (when feasible); **AND**
2. Baseline measures of outcomes also be used as controls in the impact analysis.

(RCTs with low attrition may be considered for a high rating if they meet the first of these criteria ***OR*** if they control for all three types of variables [baseline measures of outcomes, race/ethnicity, and SES]; Table 1 describes other criteria for high and moderate ratings.) Equivalence between the treatment and comparison groups must be established at baseline, that is, before the intervention being studied is provided to the program group. Establishing baseline equivalence supports conclusions that the treatment, rather than pre-existing differences, led to any observed difference in outcomes (Shadish et al. 2002). This equivalence is established if there are no statistically significant differences ($\alpha = 0.05$) on the specified variables at baseline. Equivalence must be established on the sample used in the analysis. If baseline equivalence is established on a subgroup, the results must be replicated in the same domain in two or more studies with a high or moderate rating using non-overlapping analytic study samples for the results to be considered for the determination of whether a model meets the HHS criteria (see HHS Criteria for Evidence-Based Models).

We require baseline equivalence on the demographic characteristics because they may be related to the outcome domains that are the focus of the HomVEE review. Research links SES and outcomes such as child health and child cognitive and social-emotional development (Bradley and Corwyn 2002). Similarly, outcomes may vary by the race/ethnicity of the participant. For example, research shows that birth outcomes are significantly different between different race/ethnicity groups (MacDorman 2011).

SES can be measured in multiple ways, but for this review we prefer to see equivalence on specific economic well-being measures—income, earnings, or poverty levels according to federal thresholds—because of the body of research that shows their association with child well-being, such as cognitive ability and achievement (for example, Duncan and Brooks-Gunn 1997, Fagan and Lee 2012). We also accept mean-tested assistance (such as AFDC/TANF or food stamps/SNAP receipt), maternal education, and employment of at least one member in the household if at least two such alternative measures of SES are provided, because they are closely tied to the HomVEE preferred measures of SES (income, earnings, poverty level). These measures are commonly used indicators of SES and are relevant to the population targeted to home visiting programs. In contexts outside of the nation of the United States, other measures of economic well-being will be considered.

The HomVEE review also requires that all RCTs with high attrition and all matched comparison group design studies establish baseline equivalence on the same outcomes used to examine impacts at follow-up when possible to collect baseline data on those measures. For the HomVEE review, however, researchers in some situations could not examine the same variables at baseline and follow-up. For example, if an outcome of interest is children's cognitive development, a study cannot collect baseline cognitive skills when program services start prenatally. Therefore, the variables that must be used to establish equivalence depend on whether (1) it is possible to collect the measure at baseline vs. (2) it is difficult or impossible to collect the measure at baseline. We present our criteria for these two scenarios below.

Measures assessing variables identical or sufficiently similar to the outcomes of interest are feasible at baseline. When possible, baseline equivalence should be established on the outcomes of interest; additionally, the baseline outcome should be used as a control in the analysis of impacts. Controlling for the baseline outcome ensures that any marginal differences did not bias the impact estimates at follow up. In addition, as described above, we also require that studies establish equivalence on (1) the parent's or child's race/ethnicity and (2) socioeconomic status (SES).

Measures of outcomes of interest could not be collected at baseline. For some outcomes, it is not feasible to collect baseline measures on the outcomes of interest, for example, children's cognitive and behavioral outcomes when the baseline is conducted prenatally, or parenting outcomes when parents enroll in the study before their child is born. For these studies, baseline equivalence must be established on two demographic factors, as described above: (1) the parent's or child's race/ethnicity and (2) SES.

In addition to these requirements, project leadership has the discretion to determine other cases where baseline equivalence is insufficiently demonstrated. For example, some measures that combine a wide range of responses (such as all non-White persons) may be inappropriate. In addition, a study may present comparisons for other factors at baseline, such as family structure or maternal behaviors, which are not required to establish baseline equivalence for the purposes of the HomVEE review. If a study shows statistically significant differences on these variables, it may be downgraded (that is, no longer eligible to receive the highest rating for its design). The decision to downgrade depends on the magnitude of these differences and the variables under consideration. Project leadership makes the decision in these cases, and the rationale is thoroughly documented.

Although random assignment is expected to produce groups that are equivalent, on average, on measured and unmeasured characteristics, studies with this design that otherwise meet the criteria for the high rating occasionally show statistically significant differences on selected variables (that is, race and ethnicity, SES, or an outcome measured at baseline). If such studies show that the treatment and comparison groups are not equivalent, or if it cannot be determined that the groups are similar on those factors, the variable(s)

must be used as a control in the analysis of effects. The highest HomVEE rating that random assignment studies that do not control for statistically significant baseline differences can receive is moderate.

4. Reassignment

In random assignment studies, deviation from the original random assignment can also bias the impact estimates. For example, consider a study in which a program administrator reassigned families she felt could greatly benefit from the intervention from the comparison to the intervention group. Such nonrandom selection could lead to bias in the treatment effect estimates or compromise baseline equivalence (Gartin 1995). Therefore, in order for a RCT to meet our criteria for the high rating, the analysis must be performed on the sample as originally assigned. Subjects may not be reassigned for reasons such as contamination, noncompliance, or level of exposure. RCTs that somehow alter the original random assignment but otherwise meet the criteria for the high rating are considered for a moderate study rating, provided they meet the other criteria for that rating. Our criteria are similar to those developed by the WWC, which allows a study to be downgraded as a result of reassignment.

5. Confounding factors

In certain cases, a component of the research design or methods lines up exactly with the intervention being tested, making it impossible to attribute an observed effect solely to the intervention. For example, if there is only one subject or group in the treatment or control condition, there is no way to distinguish the effects of the model from the influence of the characteristics of that one subject or group. This would occur if one home visitor were assigned to all of the families in one of the study conditions. In this case, the effect of the particular home visitor could not be separated from the treatment effect. A confounding factor could also arise from systematic differences in the way data are collected from the treatment and comparison groups—for example, if program staff collected data from all subjects in the treatment group but an independent group of staff collected data from the control group. Because the effect of the confounding factor cannot be separated from the effect of the intervention, the study findings cannot be attributed to the intervention alone (Leon 1993).

Given the severe effect that such confounding factors can have on the quality of a study, studies receive a low rating when there is either (1) only one subject or group in the treatment and control condition or (2) a systematic difference in data collection procedures between the treatment and control groups. If during the review process we find examples of other confounding factors that line up exactly with the intervention, project leadership will decide if the confounding factors may affect the study rating.

6. Clustering

If the unit of assignment is different from the unit of analysis, the analysis must account for this clustering. If a correction is not made, the statistical significance of the findings may be overstated. That is, a finding may be misclassified as statistically significant may not be when properly adjusted. If the authors do not correct for clustering at the unit of assignment, HomVEE will make an adjustment if enough information is available. The default intraclass correlations used for these corrections is 0.10, which is based on a summary of behavioral and attitudinal outcomes (WWC Procedures and Standards Handbook v2.1 – Appendix C. Clustering Correction of the Statistical Significance of Effects Estimated with Mismatched Analyses). If HomVEE does not have enough information to make the correction, the uncorrected outcomes will be excluded from the review.

C. Describing effects

Four categories are used to describe the current research findings within a HomVEE outcome domain for a selected home visiting model. The categories take into account all studies of a model that meet the HomVEE standards for a high or moderate rating. The four categories are as follows:

Favorable. A statistically significant impact on an outcome measure in a direction that is beneficial for children and/or parents. An impact could be statistically positive or negative, and it is determined "favorable" based on the end result. For example, a favorable impact could be an increase in children's vocabulary or daily reading to children by parents, or a reduction in harsh parenting practices or maternal depression. These results are represented in the tables throughout the HomVEE website in green font.

No effect. Findings for a model that are not statistically significant. These results are represented in the tables throughout the HomVEE website in black font.

Unfavorable or ambiguous. A statistically significant impact on an outcome measure in a direction that may indicate potential harm to children and/or parents. An impact could statistically be positive or negative, and it is determined "unfavorable or ambiguous" based on the end result. While some outcomes are clearly unfavorable, for other outcomes it is not as clear which direction is desirable. For example, an increase in children's behavior problems is clearly unfavorable, while an increase in number of days mothers are hospitalized is more ambiguous. This may be viewed as an unfavorable impact because it indicates that mothers have more health problems, but it could also indicate that mothers have increased access to needed health care due to their participation in a home visiting program. These results are represented in the tables throughout the HomVEE website in red font.

Not measured. Current research (meeting HomVEE standards for a high or moderate rating) does not include any measures in this domain.

HHS Criteria for Evidence-Based Models

To meet HHS' criteria 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 must either (1) be found in 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 statute, 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.[5]

[5] These criteria are consistent with the MIECHV authorizing statute: Section 511 (d)(3)(A)(i)(I)

For results from single-case designs to be considered towards the HHS criteria, additional requirements must be met:

- At least five studies examining the intervention meet the WWC's pilot single-case design standards without reservations or standards with reservations (equivalent to a high or moderate rating in HomVEE, respectively).
- The single-case designs are conducted by at least three research teams with no overlapping authorship at three institutions.
- The combined number of cases is at least 20.

Assessing Evidence of Effectiveness

After completing all impact study reviews for a model, the HomVEE team evaluates the evidence across all studies of the model that received a high or moderate rating and measured outcomes in at least one of the eligible outcome domains. The review includes outcomes for pregnant women or families with children ages birth through 5. If the sample included children outside of that age range, only subgroup analyses of children or families with children in the targeted age range are used in the review. 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 examines other aspects of the evidence for each model, including the following:

- **Quality of outcome measures.** HomVEE classifies outcome measures as primary if data were collected through direct observation, direct assessment, or administrative records; or if self-reported data were collected using a standardized (normed) instrument. Other self-reported measures are classified as secondary.
- **Duration of impacts.** To provide information on the length of follow-up, HomVEE notes when the outcomes were measured.
- **Sustained impacts.** HomVEE classifies impacts as sustained if they were measured at least one year after program enrollment.
- **Replication of impacts.** HomVEE classifies impacts as replicated if favorable, statistically significant impacts were shown in the same outcome domain in at least two non-overlapping analytic study samples.
- **Subgroup findings.** HomVEE reports subgroup findings if such findings are replicated in the same outcome domain in at least two studies using different analytic samples.
- **Unfavorable impacts.** In addition to favorable impacts, HomVEE reports unfavorable, statistically significant impacts on full sample and subgroup findings. While some outcomes are clearly unfavorable (such as an increase in children's behavior problems), others are ambiguous. For example, an increase in the number of days mothers are hospitalized could indicate an increase in health problems or increased access to needed health care due to participation in a home visiting program.
- **Evaluator independence.** HomVEE reports the funding source for each study and whether any of the study authors were model developers.
- **Magnitude of impacts.** HomVEE reports effect sizes when possible, either those calculated by the study authors or computed by HomVEE.

Implementation Reviews

The HomVEE team collects information about implementation of the prioritized models from all impact studies with a high or moderate rating and from stand-alone implementation studies. In addition, staff conduct Internet searches to find implementation materials and guidance available from home visiting model developers and national model offices. The HomVEE team uses this information to develop detailed implementation profiles (see https://homvee.acf.hhs.gov/implementation) for each prioritized model. The profiles include an overview of the model and information about prerequisites for implementation, materials and forms, estimated costs, and model contact information. National model offices are invited to review and comment on the profiles before their release. For models that meet HHS criteria for an evidence-based home visiting model, the team also extracts and reports information about implementation experiences from the studies being reviewed, including the characteristics of program participants, location and setting, staffing and supervision, model components, model adaptations or enhancements, dosage, fidelity measurement, costs, and lessons learned.

Addressing Staff Conflicts of Interest

All members of the HomVEE team who participate in prioritizing models and reviewing research sign a conflict of interest statement in which they declare any financial or personal connections to developers or products being reviewed and state their understanding of the process by which they must inform the project director if such conflicts arise. The HomVEE review team's project director assembles signed conflict of interest forms for all project staff and subcontractors and monitors for possible conflicts over time. If a team member is found to have a potential conflict of interest concerning a particular home visiting model that is being reviewed, that team member is excluded from the review process for the studies of that model. In addition, first reviews for two models evaluated by Mathematica are conducted by contracted reviewers who are not Mathematica employees.

Requests for Reconsideration of Evidence Determinations

If a State, researcher, model developer, or other interested individual believes the application of the HHS criteria for evidence of effectiveness for a particular model contains one or more errors and that, if these errors are addressed, the model would meet the evidence criteria, those concerns should be submitted to homvee@acf.hhs.gov. Inquiries are accepted through this e-mail address. Individuals may request reconsideration of the evidence-based determination based on misapplication of the HHS criteria, or missing information, or errors on the HomVEE website. HHS considers the request and if approval is granted, to ensure independence from the original review, a re-review team composed of members external to the original contractor conducts the new, independent review. The re-review team provides assurance that they do not have any actual or perceived conflicts of interest. This re-review team does not consist of members who were involved in the original review.

As with the original review, the re-review team is certified and trained in the HomVEE standards. The re-review team uses the original empirical articles (see the model reports at https://homvee.acf.hhs.gov/effectiveness), any information submitted by the individual raising the concern, and the original review team's reports, and makes any needed queries to the original team. The goal is to issue a final decision as to whether the standards were accurately applied or not within 60 days of the submission of the request for review. Following the decision, the requester is notified of the decision and HomVEE makes any necessary adjustments to the model reports or HomVEE website.

References

Bradley, R. H., and R. F. Corwyn. "Socioeconomic Status and Child Development." *Annual Review of Psychology*, vol. 53, no. 1, 2002, pp. 371–399.

Duncan, G. J., and J. Brooks-Gunn, Eds. *Consequences of Growing up Poor*. New York, Russell Sage Foundation, 1997.

Fagan, J., and Y. Lee. "Effects of Fathers' and Mothers' Cognitive Stimulation and Household Income on Toddlers' Cognition: Variations by Family Structure and Child Risk." *Fathering*, vol. 10, 2012, pp. 14–158.

Gartin, P. R. "Dealing with Design Failures in Randomized Field Experiments: Analytic Issues Regarding the Evaluation of Treatment Effects." *Journal of Research in Crime and Delinquency*, vol. 32, no. 4, 1995, p. 425.

Leon, DA. "Failed or Misleading Adjustment for Confounding." *The Lancet*, vol. 342, no. 8869, 1993, pp. 479–481.

MacDorman, M.F. "Race and Ethnic Disparities in Fetal Mortality, Preterm Birth, and Infant Mortality in the United States: An Overview." *Seminars in Perinatology (Science Direct)*, August 2011, vol. 34, no. 4, pp. 200–208.

Rubin, D. B. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine*, 1997, vol. 127, no. 8, Part 2, pp. 757–763.

Shadish, W. R., T. D. Cook, and D. T. Campbell. "Experimental and Quasi-Experimental Designs for Generalized Causal Inference." New York: Houghton Mifflin Company, 2002.

Glossary of Terms

A

Attrition. The loss of sample members from the study. Attrition typically occurs in one of three ways: (1) some sample members refuse to participate; (2) researchers are unable to locate some sample members (for example, if they have moved); or (3) researchers exclude sample members from the study (although this may negatively affect the research design). Researchers may exclude sample members for various reasons, for example, if a sample member was determined to be ineligible for the program or did not have data for all the required outcomes.

B

Baseline equivalence. The program and comparison groups have similar characteristics (such as race and education) at the study's onset. For example, the sample has baseline equivalence on language if, at the beginning of the study, similar proportions of the program and comparison groups are native English speakers and non-native English speakers.

C

Comparison group. A group with characteristics similar to those of program group members, except that those in the comparison group do not receive the services of interest. The comparison group is intended to represent what would have happened to members of the program group if they had not received the services from the model of interest. The more similar a comparison group is to the program group, the more likely it is that any difference in outcomes between the two groups can be attributed to the program.

Confounding factor. Confounding factors occur when an aspect of the study design, other than the model of interest, is conflated with the treatment or comparison group, making it impossible to measure unbiased impact. For example, if one home visitor administers all program services, it is impossible to distinguish the effectiveness of that particular person from the effectiveness of the program. Confounding factors may also arise from systematic differences in the way data are collected from subjects in the treatment group versus the comparison group. For example, participants may report information differently to someone they know, like their home visitor, than to someone they do not know, like a research assistant. Familiarity with the data collector may change the way participants answer the questions. The presence of confounding factors can significantly impede the ability of a study to capture unbiased impacts.

E

Effect size. A measure of the magnitude of the difference between the program group and the comparison group. The 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. A benefit of using the effect size is that it allows for comparisons of impacts across outcomes that may have been measured using different units. In the HomVEE review, a negative value indicates that the comparison group (which did not receive the services or program) had larger outcomes, on average, than the program group (which did receive services). A positive value indicates that the outcomes for the program group were greater than

those for the comparison group. Values of 0 indicate there is no difference, on average, between the program and comparison groups.

Evidence-based home visiting model. A home visiting model reviewed by HomVEE and found to meet HHS criteria for an "evidence-based early childhood home visiting service delivery model."

F

Favorable impact. A statistically significant impact on an outcome measure in a direction that is beneficial for children and parents. This impact could statistically be positive or negative, and is determined "favorable" based on the end result. For example, a favorable impact could be an increase in children's vocabulary or daily reading to children by parents, or a reduction in harsh parenting practices or maternal depression.

M

Mean. A measure of the average value for a sample, which equals the sum of all values divided by the number of sample members.

O

Outcome domain. A group of related outcomes that measure the same or similar constructs. The HomVEE review includes eight outcome domains: maternal health; child health; child development and school readiness; reductions in child maltreatment; reductions in juvenile delinquency, family violence, and crime; positive parenting practices; family economic self-sufficiency; as well as linkages and referrals.

P

***p*-value.** The probability that the observed finding was obtained by chance when there is no true relationship in the population. For example, a sample may show a positive mean difference, suggesting that the program group has better outcomes than the comparison group, with a *p*-value of 0.05. The *p*-value means that there is a 5 percent chance that the positive finding for the program group was obtained by chance and does not occur in the population.

Primary outcome measure. For the HomVEE review, an outcome measured through direct observation, direct assessment, or administrative data; or self-reported data collected using a standardized (normed) instrument.

Program group. The sample members who receive the services or program of interest. For the HomVEE review, the services of interest are either home visiting services or the program enhancement being tested.

Q

Quasi-experimental design. A study design in which sample members (children, parents, or families) are selected for the program and comparison conditions in a nonrandom way.

R

Randomized controlled trial. A study design in which sample members (children, parents, or families) are assigned to the program and comparison groups by chance.

Regression discontinuity design. A design in which a continuous scoring variable is used to assign an intervention to study units. Units with scores below a pre-set cutoff value are assigned to the treatment group and units with scores above the cutoff value are assigned to the comparison group, or vice versa. The effect of the intervention is estimated as the difference in mean outcomes between treatment and comparison group units, adjusting statistically for the relationship between the outcomes and the variable used to assign units to the intervention, typically referred to as the "forcing" variable.

Replicated. For the HomVEE review, favorable impacts on at least one outcome measure in the same outcome domain in at least two high- or moderate- quality studies based on different samples.

S

Sample. Persons (children, parents, or families) included in the study. For the HomVEE review, sites or cohorts that are analyzed separately are counted as separate samples.

Secondary outcome measure. For the HomVEE review, most self-reported data, excluding self-reports based on a standardized (normed) instrument.

Single-case design. These designs often involve repeated, systematic measurement of a dependent variable before, during, and after the active manipulation of an independent variable (the intervention). These designs can provide a strong basis for establishing causal inference and are widely used in applied and clinical disciplines in psychology and education.

Standardized (normed) instrument. An outcome measure that uses a uniform or standard set of procedures for administration and scoring. A norming sample, selected to be representative of the population of interest, was used to establish the standardized scoring system, or norms, for the measure.

Statistical significance. An indication of the probability that the observed finding was obtained by chance (when there is not a real relationship in the population). HomVEE uses a cutoff value of 0.05 to determine whether a finding is statistically significant.

Sustained. For the HomVEE review, favorable impacts on primary outcome measures measured at least one year after program services ended.

U

Unfavorable or ambiguous impact. A statistically significant impact on an outcome measure in a direction that may indicate potential harm to children and/or parents. This impact could statistically be positive or negative, and is determined "unfavorable or ambiguous" based on the end result. While some outcomes are clearly unfavorable, for other outcomes it is not as clear which direction is desirable. For example, an increase in children's behavior problems is clearly unfavorable, while an increase in number of days mothers are hospitalized is more ambiguous. This may be viewed as an unfavorable impact because it indicates that mothers have more health problems, but it could also indicate that mothers have increased access to needed health care due to their participation in a home visiting program.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org