Home Visiting Evidence of Effectiveness (HomVEE) Systematic Review

Handbook of Procedures and Evidence Standards: Version 2

December 2020
Emily Sama-Miller, Julieta Lugo-Gil, Jessica Harding, Lauren Akers, and Rebecca Coughlin

OPRE Report 2020-151

This page has been left blank for double-sided copying.

Home Visiting Evidence of Effectiveness (HomVEE) Handbook of Procedures and Evidence Standards: Version 2

OPRE Report 2020-151

December 2020

Emily Sama-Miller, Julieta Lugo-Gil, Jessica Harding, Lauren Akers, and Rebecca Coughlin, Mathematica

Submitted to:

Amanda Clincy Coleman and Shirley Adelstein, Project Officers
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

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

Project Director: Emily Sama-Miller
Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221

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 www.acf.hhs.gov/opre.

This page has been left blank for double-sided copying.

Acknowledgments

Public and private sector colleagues and experts made significant contributions to these revised standards. First, we acknowledge the valued support of staff at the Administration for Children and Families, U.S. Department of Health and Human Services (HHS). We particularly thank our project officers, Amanda Coleman and Shirley Adelstein, for their oversight and guidance throughout the process of developing the revised standards. We also thank Maria Woolverton, Nancy Geyelin Margie, and Naomi Goldstein for reviewing and providing thoughtful comments throughout the standards revision process. And, we appreciate the input of Jason Leger at the Maternal and Child Health Bureau in HHS's Health Resources and Services Administration.

We acknowledge the support we received from many of our colleagues at Mathematica. We particularly thank John Deke, Mariel Finucane, Peter Schochet, Elias Walsh, and Dmitriy Poznyak for discussing with us various options for new standards in detail. Sally Atkins-Burnett, Brian Goesling, and Joe Zickafoose provided valuable subject matter expertise about when and under what circumstances researchers might be expected to assess various outcome measures. We also thank Elias Walsh, Jaime Thomas, and Cay Bradley for their careful review of the resulting documentation, Larisa Wiseman, Effie Metropoulos, and John Kennedy for editing the document, Sharon Clark, Dorothy Bellow, Jessica Coldren, Dawn Patterson, and William Garrett for careful attention to layout and production, Brigitte Tran, Laura Watson-Sarnoski, and Sheena Flowers for graphic design, and Katherine Niland and April Wu for assistance with handling revisions.

In addition, we are grateful to the experts who reviewed and consulted on the handbook; their thoughtful feedback helped us to improve its contents. The views expressed in this publication do not necessarily reflect the views of these experts.

Sandra Wilson
Abt Associates

Jill Filene
James Bell Associates

David MacKinnon
Arizona State University

Brenda Jones Harden
University of Maryland

Elizabeth Stuart
The Johns Hopkins University

This page has been left blank for double-sided copying.

Contents

This page has been left blank for double-sided copying.

Exhibits

I. Introduction

This handbook describes the methods used by the Home Visiting Evidence of Effectiveness (HomVEE) review to review existing research and report the findings. It is designed for use by researchers, policymakers, and other stakeholders.

A. Background

Home visiting is increasingly used to deliver services to at-risk families with young children. All 50 states and the District of Columbia have home visiting programs (Stoltzfus and Lynch 2009; National Home Visiting Resource Center [NHVRC] 2018). In 2019 alone, nearly 300,000 families received more than 3 million home visits from evidence-based models, and millions more families are poised to benefit from this type of service (NHVRC 2019). As home visiting models become more widespread, there is increased interest in offering models that have established evidence of effectiveness.

The mission of the HomVEE review is to conduct a thorough and transparent review of early childhood home visiting models. HomVEE provides an assessment of the evidence of effectiveness for early childhood home visiting models that serve families with pregnant women and children from birth to kindergarten entry.

HomVEE assesses the quality of the research evidence; not all evidence is based on equally well-designed research. Systematic reviews, such as HomVEE, methodically select a pool of research to review, identify well-designed research within that pool, and then extract and summarize the findings from that research. HomVEE's work helps policymakers and program administrators understand which models are effective. It is important to note that HomVEE does not directly evaluate home visiting models. Instead, it reviews and reports on the findings of existing research that does evaluate them. Specifically, HomVEE focuses on reviewing research that examines early childhood home visiting models (see Exhibit I.1). The HomVEE review was launched in 2009 with sponsorship from the Administration for Children and Families (ACF) Office of Planning, Research, and Evaluation (OPRE) within the U.S. Department of Health and Human Services (HHS).

Exhibit I.1. HomVEE's definition of an early childhood home visiting model

HomVEE defines an **early childhood home visiting model** as an intervention in which trained home visitors meet with expectant parents or families with young children to deliver a specified set of services through a specified set of interactions. These programs are voluntary interventions that are either designed or adapted and tested for delivery in the home. During the visits, home visitors aim to build strong, positive relationships with families to improve child and family outcomes. Services may be delivered on a schedule that is defined or can be tailored to meet family needs. A model has a set of fidelity standards that describe how the model is to be implemented.

Models reviewed by HomVEE must serve pregnant women or families with children from birth to kindergarten entry (that is, through age 5), and the primary service delivery strategy must be home visiting. In addition, the model must have research that examines its effects in at least one of eight 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.*

**Note: These domains are inclusive of the benchmark domains and individual outcomes listed in the statute that authorized the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) Program (Social Security Act, Section 511 [42 U.S.C. 711])*

One critical use of HomVEE's results is to determine which home visiting models meet the HHS criteria for an "evidence-based early childhood home visiting service delivery model," a key requirement of eligibility for programs implemented with funding from the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) Program. For the purposes of the HomVEE review, this handbook uses the term "evidence-based model" to refer specifically to a model that meets HHS criteria developed based on statutory requirements in the authorizing legislation for the MIECHV Program. HomVEE recognizes that other systematic reviews may use different criteria to evaluate evidence of effectiveness. Thus, an evidence-based model in the context of HomVEE might or might not meet requirements for evidence of effectiveness according to other systematic reviews.

Created in 2010, the MIECHV Program provides funding to states, territories, and tribal entities to implement home visiting models. MIECHV awardees have the flexibility to tailor their programs to serve the specific needs of their communities. They perform a needs assessment to identify at-risk communities and select the best home visiting service delivery models for their state and/or local needs. As per MIECHV's authorizing statute, state and territory awardees must spend the majority of their MIECHV Program grants to implement evidence-based home visiting models, with up to 25 percent of funding available to implement promising approaches that will undergo rigorous evaluation. In accordance with the flexibility provided by the MIECHV authorizing statute for grants to tribal organizations, Tribal MIECHV grantees can use up to 100 percent of their MIECHV grants for promising approaches that will undergo rigorous evaluation.

The MIECHV Program is administered by the Health Resources and Services Administration (HRSA) in partnership with ACF. In February 2018, the MIECHV Program was allocated $400 million per year through fiscal year (FY) 2022. In 2019, MIECHV-funded programs alone served about 79,000 families (HHS 2020), and states and localities often offer non-MIECHV funded home visiting services to additional families as well (NHVRC 2018). A HomVEE designation as an evidence-based model does not guarantee that a model is eligible to be implemented with MIECHV funding. To be eligible for

implementation as an evidence-based model with MIECHV funding, a model must both meet HHS criteria for evidence of effectiveness (as determined by HomVEE) and meet all other statutory requirements for model eligibility (as determined by HRSA). In addition, MIECHV's authorizing statute allows awardees to utilize a portion of their MIECVH funding for a model that qualifies as a promising approach.[1]

For the first time since its inception, HomVEE has substantially revised several procedures and standards for its review (see Exhibit I.2). Notably, HomVEE has now defined the term "early childhood home visiting model" in Exhibit I.1. HomVEE consulted with selected methods and content experts outside of the HomVEE contractor, including with other federal evidence reviews to refine and update the procedures and standards in this document. These changes bring HomVEE generally into alignment with procedures and standards for other federally sponsored systematic evidence reviews, and address critical topics in the evolution of the home visiting field. HomVEE also consulted with and an HHS work group that collaboratively developed the new definitions and rules. An earlier draft of this handbook was released for public comment in August 2020. This final HomVEE Version 2 handbook adjusts that earlier draft to respond to public and expert consultants' feedback.

HomVEE will apply these procedures and standards beginning with the 2021 annual review. In addition, HomVEE will retroactively apply its clarified *terminology* and certain procedures. Specifically, to promote consistency in reporting across the review, clarifications about the outcomes eligible for review in each domain and the clarified definitions of study, manuscript, and subgroup retroactively will apply to all research on models regardless of (1) the model's evidence-based status according to HHS criteria, (2) whether the model is prioritized and selected for review, and (3) whether HomVEE previously reviewed the manuscript.

HomVEE generally will not retroactively apply the new *standards* to previously reviewed research about evidence-based models unless it is single-case design research about a model HomVEE selects for review. For example, manuscripts that have ineligible findings of findings moved to other outcome domains based on retroactive application of clarified terminology and procedures will not be reviewed again with HomVEE Version 2 standards.

Exhibit I.2. Summary of updates to HomVEE procedures and standards in Version 2 Handbook

	Topic	Description of change or clarification	Where to find out more
1	Clarify definitions of research terms	Defined key terms, including early childhood home visiting model; and study, manuscript, subgroup, and finding. This clarification recognizes the importance of clear communication and consistent terminology when applying systematic review rules.	Exhibit I.1, Exhibit I.4,

[1] For additional information on the MIECHV Program, see https://mchb.hrsa.gov/maternal-child-health-initiatives/home-visiting-overview.

	Topic	Description of change or clarification	Where to find out more
2	Changes to terminology used	Previously, HomVEE used the term "matched comparison group design" to refer to quasi-experimental designs (QED) with nonrandomly formed comparison groups. HomVEE will now use the term non-experimental comparison group design (NED) for this same category or research. This change clarifies that this category of QEDs includes both comparison groups formed through matching and other types of nonrandom group designs, such as convenience samples. HomVEE will continue to use QED as an umbrella term including regression discontinuity design, single-case design, and NED research.	Chapter I, Section C, or Chapter III, Section A.1
3	Establish a 20-year moving search window for reviewing most manuscripts	HomVEE implemented a 20-year moving window for previously unreviewed manuscripts to be eligible for review, with an exception for manuscripts that have been or are submitted to the annual call for research. This change keeps the review focused on more current research.	Chapter II, Section A.1.a.1
4	Implemented the PRESS method for systematic searching to revise the search terms for the annual review	HomVEE uses a modified Peer Review of Electronic Search Strategies (PRESS) method to refine the search terms (McGowan et al. 2016). With this method, a pair of trained librarians use a structured tool to map the search terms to the systematic review scope to enhance the quality and comprehensiveness of the search. This change recognizes accepted practice in the library science field.	Chapter II, Section A.1.a.1
5	Add new grey literature databases	HomVEE has expanded its annual search to include four additional databases to identify manuscripts that are not published in journals, or grey literature. These databases are Google Scholar, the Harvard Think Tank Search, and a pair of preprint servers (the Open Science Framework database and MedRxiv). This change helps HomVEE to confirm that its literature search approach is open and comprehensive.	Chapter II, Section A.1.a.1
6	Changes to prioritization point allocation (manuscript level)	HomVEE has adjusted the relative point value that it assigns to each manuscript for populations served by home visiting models, including those named as priority populations in the MIECHV statue. The changes are in response to stakeholders' feedback submitted during the public comment period that emphasized the importance of carefully considering which specific populations should get points from HomVEE at this stage.	Chapter II, Section A.2.a.

	Topic	Description of change or clarification	Where to find out more
7	Establish rule for accepting supplemental information	Supplemental information can take two forms: (1) **new information** about a study's methods or procedures; or (2) **new research** that supplements what HomVEE had on hand at the close of that year's call for research, such as additional findings or new analyses of research in a previously reviewed manuscript, or an entirely new set of findings. HomVEE has clarified its approach to how the team incorporates information received from study authors when HomVEE reviews manuscripts. Previously, HomVEE had no published procedures on this topic and the clarification will support consistent handling of supplemental information across the review.	Chapter II, Section B.1.b, Exhibit II.9
8	Defining subgroups and protocol for reporting subgroup analyses	HomVEE defines a subgroup as a subset of the *overall sample examined in a study*—that is, an **analytic subgroup.** Notably, this differs from defining subgroup as a subset of the *overall population.* In reviews of manuscripts about well-designed impact studies, HomVEE lists the subgroups examined by research. (*Details* on the findings will be listed only for replicated subgroups.) This definition, and focus on reporting replicated subgroups, clarifies how HomVEE applies HHS' criteria for an "evidence-based early childhood home visiting service delivery model" with respect to subgroup analyses.	Exhibit I.4, Exhibit II.10
9	Contrasts in impact research that are ineligible for review by HomVEE	HomVEE generally excludes research that isolates the impact of model features. This clarification is to emphasize that such research does not answer HomVEE's core question for the annual review of whether an early childhood home visiting model is effective. (However, HomVEE recognizes that research on specific features is potentially useful for many other purposes).	Chapter III, Section A.2
10	Ineligible and preferred analyses	HomVEE excludes certain analyses within manuscripts about impact studies as ineligible. This clarification is intended to emphasize that questions about the mechanisms behind how a model works, the settings where it might work best, and the populations who benefit the most from the intervention are outside of the scope of the HomVEE annual review.	Chapter III, Section A.3

	Topic	Description of change or clarification	Where to find out more
11	Eligible outcomes and baseline assessability	HomVEE has clarified: • That it reports only unique findings (those that report results on a different outcome, sample or subgroup, or time period, or with a different analytic approach, from findings reported in other manuscripts about the same home visiting model) • Which findings are eligible for review in each domain • Which outcomes are assessable at baseline, under what circumstances These clarifications formalize and expand HomVEE's existing internal guidance to reviewers.	Chapter III, Section A.4, Exhibit III.2, and Appendix B
12	Some sample loss does not count as attrition	Some types of sample loss do not count as attrition in HomVEE, including losing sample due to acts of nature and randomly selected subsamples. This change aligns with Version 4.1 of the What Works Clearinghouse (WWC) Standards.	Chapter III, Section B.1
13	No baseline equivalence requirement for low-attrition randomized controlled trials (RCTs)	HomVEE no longer requires that authors of RCTs with low attrition establish equivalence nor that they adjust for baseline differences. This change aligns with Version 4.1 of the WWC Standards.	Chapter III, Section B.2
14	Baseline equivalence depends on difference in effect sizes and other considerations	HomVEE assesses baseline equivalence based on the magnitude of the difference in standard deviation units (effect size). This is a change from original HomVEE standards that assessed equivalence based on statistical significance. This change aligns with Version 4.1 of the WWC Standards. In addition, HomVEE now considers maternal education to be a primary measure of socioeconomic status (SES) at baseline (rather than a secondary measure that requires equivalence on another SES measure as reinforcement.) This change is based on subject matter expert feedback and a consultation of research on this topic.	Chapter III, Section B.2.b
15	Allowable statistical adjustment techniques	HomVEE has indicated which statistical adjustment procedures are acceptable. This list follows WWC Version 4.1 guidelines, and notes a few additional statistical adjustment procedures that HomVEE reviewers will accept (especially for research designs such as repeated measures analyses for which HomVEE has specific standards that WWC Version 4.1 does not cover).	Chapter III, Section B.2.b

	Topic	Description of change or clarification	Where to find out more
16	Requirement for validity and reliability of outcome measure	HomVEE has established face validity and reliability standards. The standards align HomVEE with other federally sponsored systematic reviews, including WWC and the Prevention Services Clearinghouse. HomVEE reviewers apply these new standards to all findings that are within one of HomVEE's eight outcome domains and to all measures HomVEE uses to assess baseline equivalence. Findings about outcomes that do not meet both the face validity and the reliability standard rate low.	Chapter III, Section B.4
17	Cluster RCTs	A finding in a manuscript about a cluster RCT is eligible for a high rating only if it has (1) low sample loss at the cluster level, (2) a random assignment design that is not compromised, and (3) no risk of bias due to nonresponse at the family unit/individual level. To rate moderate, research about cluster RCTs with high attrition and research about cluster NEDs must demonstrate baseline equivalence of the analytic sample of family units/individuals, or either (1) have an analytic sample of family units/individuals that is representative of the clusters and have an RCT design with low sample loss at the cluster level or (2) have an analytic sample of family units/individuals that is representative of the clusters and establish equivalence of the clusters at baseline using a representative sample for groups in the analytic sample. These new standards generally align with Version 4.1 of the WWC Standards (see Chapter III, Section C.1 for details).	Chapter III, Section C.1
18	Standards for addressing missing data	HomVEE has detailed guidance about how authors must handle missing data for manuscripts about impact studies to earn a rating of at least moderate. The guidance aligns with WWC Version 4.1 standards on this point.	Chapter III, Section C.2, and Appendix E
19	Repeated measures analyses	In repeated measures analyses, researchers measure the research sample at several time points to chart its growth over the course of the intervention and, sometimes, beyond. HomVEE only reviews and report findings from repeated measures analyses with multiple follow-ups in RCTs and NEDs when impact estimates and details about statistical significance are available, in the manuscript or from authors, for individual time points. Research with this design is becoming increasingly common in the home visiting evaluation field. Earlier, HomVEE had not specified any standards for reviewing repeated measures analyses in group-design studies (such as RCTs and NEDs), nor have other federally sponsored systematic evidence reviews thoroughly addressed this.	Chapter III, Section C.2

	Topic	Description of change or clarification	Where to find out more
20	Structural equation models (SEMs)	SEMs are a statistical modeling technique that analyzes the structural relationships between variables, often including both observed and unobserved or latent constructs. SEMs are eligible for review only if they are accompanied by a path diagram (including one authors may submit in response to a query from HomVEE) and that are identified (that is, the degrees of freedom are greater than the parameters to be estimated). When reviewing an SEM diagram to identify which outcomes to review, HomVEE reviewers ask: In the SEM diagram, is there a direct pathway from the intervention to the outcome? and: Are there no pathways leading *to* that outcome *from* another outcome? If both answers are yes, that outcome is eligible for review. Research with this design is becoming increasingly common in the home visiting evaluation field. Earlier, HomVEE had not specified any standards for reviewing repeated measures analyses in group-design studies (such as RCTs and NEDs), nor have other federally sponsored systematic evidence reviews thoroughly addressed this.	Chapter III, Section C.3
21	Adopt new WWC Version 4.1 standards for regression discontinuity designs (RDDs)	HomVEE has aligned its standards to WWC's latest (Version 4.1) RDD standards and procedures.	Appendix C
22	Review of single-case design (SCD) research	HomVEE instructs reviewers to calculate and use a design-comparable effect size (D-CES) to characterize the findings. Reviewers still use visual analysis to assess whether an SCD study is well designed. This change is to align with WWC's Version 4.1 procedures. (Note: A forthcoming update to the Handbook will provide additional information about how HomVEE will incorporate the D-CES into decisions about whether a model is evidence based according to HHS criteria).	Appendix D

B. Organization of this handbook

This handbook of procedures and standards is a transparent account of how the HomVEE review operates. As noted, this is the second version of HomVEE's procedures and standards, and it updates the version initially published on the HomVEE website.[2]

- This first chapter gives some background about HomVEE, including the scope of the review and definitions of key terms (Exhibits I.1 and I.4).
- Chapter II describes the evidence review process, including how HomVEE (1) identifies eligible research and prioritizes models for review, (2) rates the quality of impact research and assesses whether the model is evidence based, and (3) reports results on the model's impact and implementation.
- Chapter III describes the standards for rating research quality, and how HomVEE applies those standards to rate the quality of impact research.

Technical details about the procedures and standards are in the appendices.

C. Evidence examined by HomVEE

Systematic reviews of evidence define their scope based on population, intervention, comparators, outcomes, timing, and setting (PICOTS) (Thompson et al. 2012). Exhibit I.3 uses the PICOTS criteria to summarize the scope of the HomVEE systematic review effort.[3] Among research that fits within HomVEE's scope, the team then identifies which models are evidence based. The review also reports detailed information about the samples of families who participated in the research, the outcomes measured in each manuscript, and the implementation guidelines for each model.

Exhibit I.3. Evidence examined by HomVEE review

PICOTS criterion	HomVEE's treatment of criterion
Population	Families with pregnant women or with children from birth to kindergarten entry (through age 5)
Interventions	Given limited resources, HomVEE prioritizes certain **early childhood home visiting models** (defined in Exhibit I.1) to review**.** Research evaluating the impact of model feature(s) is generally ineligible for review because it does not answer the question of whether a multi-feature model is effective overall (see Chapter III, Section A.2 on contrasts that HomVEE reviews).
Comparators	Comparison groups that are offered services typically provided to pregnant women or families with young children, or other programs and policies for which they might be eligible.
Outcomes	Eight 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.

[2] Those original procedures and standards are now archived here: https://homvee.acf.hhs.gov/publications/methods-standards.

[3] The review plans specified here also address each section of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis for Protocols (PRISMA-P; Moher et al. 2015) and the methods section of the PRISMA for Complex Interventions (PRISMA-CI; Guise et al. 2017b). This handbook also serves as the protocol for the review (PRISMA-CI element 5). A checklist version of the PRISMA elements is in Appendix A.

PICOTS criterion	HomVEE's treatment of criterion
Timing	Analyses published or prepared in the previous 20 years. The services implemented within an intervention can be of any duration, and the outcomes can be measured at any length of follow-up. (HomVEE uses a 20-year moving window for its literature search, but considers older research for the review if stakeholders submit it or if it has previously been reviewed by HomVEE.)
Setting	Manuscripts prepared in English and describing research conducted in a developed-world context, defined as countries that had high incomes in the year the manuscript was published (or, for unpublished research, the year it was submitted to HomVEE) according to the World Bank Indicators, are generally eligible. *Exception: Research on **evidence-based** models that is conducted outside of the United States (but still in a developed-world context) is excluded unless HomVEE resources in a given year permit review of that research.*

Note: Classification based on PICOTS framework. See Thompson et al. (2012) and World Bank (2020).

HomVEE reviews manuscripts about impact studies to determine which impact studies are well designed; based on findings from well-designed impact studies that are well executed, HomVEE identifies which home visiting models are evidence based. **Well-designed impact studies** are those whose design and execution suggest that some or all of the findings were due to the home visiting model rather than other factors. Specifically, HomVEE considers two types of study designs to be reliable for answering the question of whether a home visiting model is effective: (1) randomized controlled trials (RCTs) and (2) quasi-experimental designs (QEDs). Eligible QEDs include single-case, regression discontinuity, and non-experimental comparison group designs. These designs are described in detail in Chapter III.

Notably, the main focus of HomVEE's annual review is to apply HHS' criteria for an evidence-based early childhood home visiting service delivery model (Exhibit II.11), which requires rating the quality of impact studies. To do so, because these criteria focus on findings and the research sample in which the finding is observed, HomVEE must examine each finding presented within each manuscript about the study. (A study's findings might be presented across several manuscripts.) Specifically, HomVEE assigns a rating to each manuscript based on the degree of confidence that its reported findings are a result of the home visiting model. Exhibit I.4 describes how HomVEE defines other key research terms.

HomVEE also reviews implementation research to support summaries about home visiting models, as described at the end of Chapter II.

Exhibit I.4. HomVEE's definitions of key research terms

Home visiting researchers may study the same sample over many years and report results in several places. Therefore, HomVEE relies on specific terminology to classify research:

- A **study** evaluates a distinct implementation of an intervention (that is, a home visiting model implemented with a distinct sample, enrolled into the research investigation at a defined time and place, by a specific researcher or research team). HomVEE reviews eligible manuscripts about studies that examine the impact of an early childhood home visiting model by comparing an **intervention condition** (in which study participants are offered the home visiting model under study) and a **comparison condition** (in which study participants are not offered that model). See Chapter II, Section A.1.b, including Exhibit II.4, for more information on how HomVEE screens research for eligibility.
    - A **sample** encompasses both the entire intervention group and the entire comparison group. (Note: in studies that use a single-case design, the sample participants receive both the intervention and the comparison condition.)
    - A **subgroup** is a subset of the sample examined in a study (that is, an analytic subgroup). For example, researchers may examine how a home visiting model affects teenage mothers when there are mothers with a range of ages in their study; hence, teenage mothers would be an analytic subgroup. Sometimes researchers present subgroup findings in a manuscript alongside findings for the overall sample, and sometimes researchers prepare a manuscript based exclusively on subgroup findings from a broader study. (For HomVEE, results from teenage mothers would not be considered an analytic subgroup analysis if the overall study only enrolled teenage mothers. See Chapter II, Section B.2.b for details on how HomVEE handles subgroup research.)
- **Manuscripts** describe study results. Manuscripts may be published or unpublished research, such as journal articles, book chapters, or working papers. A single study may produce one or many manuscripts. Typically, one manuscript reports on only one study, although in rare cases one manuscript may include several studies, if it describes evaluations of multiple interventions (such as multiple versions of a home visiting model) or the same intervention evaluated in multiple distinct (non-overlapping) samples (such as different cohorts over time, or in multiple, independent locations).
- **Findings** summarize the effect of a home visiting model on a specific sample or subgroup, on a specific eligible outcome measure (see Chapter III), at a specific time point, from a specific analysis. A manuscript typically includes multiple findings.

  HomVEE rates findings (see Chapter III) and sorts manuscripts according to the highest-rated finding in the manuscript (see Chapter II). When determining which models are evidence based, HomVEE considers both whether the research that calculated the findings was well designed and whether the findings come from different studies (with distinct samples). See Exhibit II.11 for details.

This page has been left blank for double sided copying.

II. Evidence Review Process

Every year, HomVEE uses standardized techniques to systematically identify and review research about home visiting models. The goal is to use findings from well-designed, well-executed studies of the impact of home visiting to identify "evidence-based early childhood home visiting service delivery model[s]" according to criteria defined by the U.S. Department of Health and Human Services (HHS).[4]

This chapter describes the HomVEE systematic review process (see Exhibit II.1) in detail, but to summarize, the process involves using independent, unbiased reviewers to:[5]

- Locate research and prioritize models for review.
- Review eligible research on prioritized models, and use HomVEE's published standards to rate the quality of the impact study described in each manuscript.
- Examine findings across all manuscripts on a study and all studies on a model and identify evidence-based models.
- Review and summarize information about how prioritized models were implemented.
- Publish results from the review in reports on model effectiveness research and implementation on the HomVEE website (https://homvee.acf.hhs.gov/).

[4] For the purposes of the HomVEE review, this handbook uses the term *evidence-based model* to refer specifically to a model that meets HHS criteria developed based on statutory requirements in the authorizing legislation for the MIECHV Program. HomVEE recognizes that other systematic reviews may use different criteria to evaluate evidence of effectiveness. Thus, an evidence-based model in the context of HomVEE might or might not meet requirements for evidence of effectiveness according to other systematic reviews.

[5] Conducting an accurate review with integrity requires that staff participating in any step of the process that could affect a decision about whether research is well designed or whether a model is effective must be free of conflicts of interest. All members of the HomVEE contractor review team who are involved with the search, screening, and review process sign a conflict of interest statement in which they declare any financial or personal connections to model developers or products being reviewed. The conflict of interest statement also outlines the process by which members of the HomVEE contractor review team must inform the project director if such conflicts of interest arise. The HomVEE contractor review team's leadership assembles signed conflict of interest forms for all relevant contractor and subcontractor staff and monitors the team for possible conflicts over time. If a team member is found to have a potential conflict of interest concerning a home visiting model under review, that team member is excluded from the review process for the studies of that model. In addition, if the HomVEE contractor conducted a study of a home visiting model being reviewed, a reviewer external to that contractor's staff conducts the review of all manuscripts related to that study, and a reviewer from the contracted firm simply confirms that the review documentation has been fully completed.

Exhibit II.1. HomVEE evidence review process

A. Prioritize

The first part of HomVEE's process involves searching and screening manuscripts, calculating prioritization scores for each eligible home visiting model identified in a search, and selecting models to review.

1. Search and screen

The HomVEE evidence review must be thorough so it identifies all models that may be evidence-based models.

From October through early January of the following year, HomVEE conducts a broad annual search for research on home visiting models serving pregnant women or families with children whose ages range from birth to kindergarten entry (that is, up through age 5) and carefully screens the resulting manuscripts for eligibility.[6]

The search aims to locate research on home visiting models that are designed to improve outcomes in at least one of the following eight domains:[7]

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
8. Reductions in juvenile delinquency, family violence, and crime

[6] The screening process described in this section refers to the annual review. HomVEE may refine the screening approach as relevant for specific stand-alone products, such as the review of research on tribal populations and the *Evidence Says* series.

[7] These domains were selected to align with the benchmark domains and participant outcomes specified in the statute authorizing the MIECHV Program (Social Security Act, Section 511 [42 U.S.C. 711]).

a. Search strategy

There are two main activities in HomVEE's annual literature search:[8]

1. **Database searches.** HomVEE searches on relevant keywords in a range of research databases. This search identifies new manuscripts that have been released from the previous October through the end of September. Keywords include terms related to interventions that are eligible for the review, population, and relevant outcome domains of interest (Exhibit II.2). HomVEE also performs focused searching, by model name, both for evidence-based models and for other models with the highest prioritization scores (see Section A.2 of this chapter about how HomVEE calculates prioritization scores).

   Searches in HomVEE's first 11 annual reviews were for manuscripts published in or after 1989. To keep the review current, HomVEE will implement a 20-year moving window for manuscripts to be eligible for review. HomVEE will implement this moving window beginning with the 2021 review (see Section 1.b, below, on screening criteria). However, two categories of older research will still be eligible for HomVEE: (1) older research that HomVEE has already reviewed, and (2) research submitted at any time (that is, since HomVEE's inception and moving forward) through the call for research, as described in the next section.

   To ensure that the search strategy is thorough, replicable, and meets the research objectives, HomVEE uses a modified Peer Review of Electronic Search Strategies (PRESS) method to refine the search terms (McGowan et al. 2016). With this method, trained librarians use a structured tool to map the search terms to the PICOTS criteria to enhance the quality and comprehensiveness of the search by checking for things such as correct use of Boolean search operators, and alternate words and spellings for search terms. One librarian carefully searches the selected electronic databases, documenting each step of the process, and another applies most of the PRESS 2015 Evidence-Based Checklist (McGowan et al. 2016) to provide guidance and check the results.[9]

Exhibit II.2. Keywords used in the database searches

Category	ID	Search term
Search restrictions	--	Manuscripts published in English only Manuscripts published within past year
Intervention	S1	(home AND visit*) or "family development" or (case AND manage*) or ((coordination OR referral*) AND (home AND visit*))
Population	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*"

[8] In addition to these two activities, in the first year of the review, HomVEE also included (1) a review of existing literature reviews and meta-analyses, to confirm that the search strategy was capturing essential research, and (2) a custom Google search engine to examine more than 50 specific, relevant websites, which was discontinued because the results largely overlapped with the results of the database searches and call for research.

[9] One step in the PRESS method involves checking each database's list of subject terms (a defined list of topics controlled by each database) and adjusting search terms to make sure differences in the subject terms are captured. However, the HomVEE search terms are designed to be broad enough to capture research regardless of how databases define their subject terms. Thus, HomVEE eliminated this step from its process.

Category	ID	Search term
Outcomes	S3	(child* and (abuse or neglect or maltreatment or health or injur* or violence or attachment or immuniz* or "emergency department*" or "emergency service*")) 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 families 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) or "positive parenting" or "family engagement" or "family involvement" or "parent-child interaction"
Document type	S4	(study or evaluat* or research) and (effective* or efficac* or impact* or outcome* or implement* or cost or replic*)
Combine terms	S5	(S1 AND S2 AND S3 AND S4)

Note: In the Version 2 Handbook, HomVEE has made minor revisions to this table to adjust truncation of some terms (for instance, to add an asterisk after "child" to capture results on child or children) that did not already have the truncation indicated. The Version 2 Handbook also updates the search period to reflect a 20-year moving window for screening database search results. To implement this, HomVEE adds the newest year of database search results with each annual cycle, and drops the oldest year of results except for previously reviewed manuscripts and submissions to the call for research.

HomVEE (including for the 2021 review) fully searches the following databases:[10]

- Academic Search Premier
- APA PsycInfo
- Campbell Collaboration*
- CINAHL with Full Text
- Cochrane Central Register of Controlled Trials
- Cochrane Methodology Register
- E-Journals
- EconLit
- Education Research Complete
- ERIC
- MEDLINE
- New York Academy of Medicine's Grey Literature Report*
- ProQuest Dissertations
- SAGE database*

[10] To ensure the review has the most useful and relevant information, HomVEE rarely drops or adds a database. When changes to the list of databases occur, HomVEE will reflect it in updates to the published handbook.

- Scopus
- SocINDEX with Full Text
- WorldCat*

Some databases do not support searching with long strings of search terms. For these databases, denoted above with an asterisk (*), HomVEE uses an abbreviated list of search terms to identify the most relevant literature.

In addition, HomVEE searches the following databases that focus on grey (that is, unpublished) literature to identify publicly available unpublished research, and to emphasize the value of open science practices. This step specifically occurs by searching the names of top-scoring models (as calculated based on the prioritization points described in the next section of the handbook). In this way, HomVEE can be confident of capturing all relevant literature about all models that have sufficient points to be considered for review in a given year.

- Child Care & Early Education Research Connections
- Google Scholar
- Harvard Think Tank Search
- MedRxiv (preprint server)
- Open Science Framework database (preprint server)

2. **Call for research.** In addition to conducting database searches, HomVEE issues an annual call for research each fall. The call is published on the HomVEE website, shared with subscribers to the HomVEE mailing list, and sent to more than 40 relevant electronic mailing lists or organizations for dissemination (Exhibit II.3). The call for research is typically released each November and is open through early January of the following year, and it accepts unpublished manuscripts and manuscripts published through December of the previous year. (If authors submit unpublished research, it should be in the form of a full manuscript with enough text describing the study's procedure, analysis approach, and findings for HomVEE to conduct its review.) Members of the public may submit manuscripts at any time during the year. However, HomVEE will only consider those submitted before the call for research closes in January for that year's review cycle. HomVEE will consider manuscripts submitted after January in a future year's review cycle.[11]

[11] Each year, HomVEE screens all submitted manuscripts for relevance and prioritizes models for review as described later in this chapter. HomVEE retains all submissions that are eligible for review, but because of the volume of research received through the call and identified through database searches, HomVEE cannot review all submitted manuscripts. In a given year, HomVEE only reviews the impact and implementation manuscripts on models prioritized for review that year. HomVEE will consider submissions that do not focus on one of the prioritized models in subsequent review cycles, and HomVEE will review those only when the model the manuscript discusses is selected for review.

Exhibit II.3. Distribution of HomVEE annual call for research

- American Academy of Pediatrics
- American Education Research Association
- American Medical Association
- American Professional Society on the Abuse of Children
- American Psychiatric Nurses Association
- American Psychological Association
- American Public Health Association
- American Sociological Association
- Association for Public Policy Analysis and Management
- Association of Black Psychologists
- Association of Maternal and Child Health Programs
- Association for Psychological Science
- Child Care and Early Education Research Connections
- Child Welfare Information Gateway
- Center for Law and Social Policy
- Collaborative for Understanding the Pedagogy of Infant/Toddler Development
- Congressional Research Service
- Early Childhood Comprehensive Systems Impact grantees
- Federal Inter-Agency Workgroup on Child Abuse & Neglect
- Foundation for Child Development
- FRIENDS electronic mailing list (for Community-Based Child Abuse Prevention grantees and interested community members)
- Harvard's Center on the Developing Child
- Home Visiting Applied Research Collaborative
- HomVEE website and mailing list
- Institute for the Advancement of Social Work Research
- International Society for the Prevention of Child Abuse and Neglect
- National Association for the Education of Young Children
- National Association for Welfare Research and Statistics
- National Association of County and City Health Officials (NACCHO)
- National Association of Social Workers
- National Council on Family Relations
- National Governors Association
- National Prevention Partners
- Network of Infant/Toddler Researchers (NITR)
- Office on Child Abuse and Neglect
- Pew Charitable Trusts
- Prevent Child Abuse America
- Society for Prevention Research
- Society for Research in Child Development
- Society for the Psychological Study of Social Issues
- Society of Pediatric Nurses
- State/territory MIECHV awardees
- State/territory MIECHV TA providers
- Tribal MIECHV grantees
- Tribal MIECHV TA providers
- W.K. Kellogg Foundation
- World Association for Infant Mental Health
- Zero to Three

b. Screening criteria

Next, HomVEE uses information from the title and abstract to screen the results of the database searches and the call for research for their relevance to and eligibility for HomVEE review.[12] When the title and abstract do not provide enough information to clearly indicate that a manuscript is not eligible for HomVEE review, HomVEE screens it in. HomVEE screens out manuscripts for any of the following reasons:[13]

[12] For models that fall into the top tier of prioritization scores in the process described in the next section, HomVEE re-screens manuscripts using information from the full text, and adjusts prioritization scores as needed.

[13] These criteria apply to both the title and abstract and full-text screening stages (full-text screening is described under Step 4 of calculating the prioritization score, in the following section). HomVEE recognizes that this information might not be available in the title and abstract alone; therefore, when it is unclear whether a manuscript is eligible for review, HomVEE will screen it in at the title and abstract stage and examine its full text more carefully to make a screening decision.

- The manuscript examines a home visiting model implemented in a mandatory setting (for example, if families are required to participate as part of a residential treatment program or a child custody agreement).
- Home visiting was not the primary service delivery strategy studied in the intervention. (For example, models that provide services primarily in centers, with supplemental home visits, are excluded.)
- The study that the manuscript examines did not use an eligible design. Eligible designs for impact studies are randomized controlled trials and three types of quasi-experimental designs: (1) single-case designs, (2) regression discontinuity designs, and (3) non-experimental comparison group designs (see Chapter III).[14]
- The manuscript did not report results for an eligible target population. Eligible target populations are pregnant women or families with children whose ages range from birth to kindergarten entry (that is, up through age 5) and who are served in a developed-world context.[15]
- The manuscript did not examine any eligible comparisons (see Chapter III, Section A.2 for information on contrasts that HomVEE reviews).
- The manuscript did not examine any findings in HomVEE's eight eligible outcome domains, listed in Exhibit I.1.
- The manuscript did not examine a home visiting intervention. (For example, the manuscript examined a grant program and its grantees, a medical intervention delivered by home nurses, or legislation.)[16]
- The manuscript was not published in English.
- The manuscript was published more than 20 years ago, unless it was submitted to the call for research or has already been reviewed by HomVEE.[17] For example, for the 2021 review, HomVEE will consider previously unreviewed manuscripts released or published from 2001 through 2020, as well as any submissions to the call for research and any previously reviewed manuscripts.

[14] HomVEE generally will not review the quality of a manuscript that isolates the effect of a model feature on child or family outcomes (see Exhibit III.1 for details). HomVEE treats manuscripts that examine implementation outcomes as implementation research, which does not contribute to a model's evidence rating. Prior to the 2021 review, HomVEE also reviewed implementation research, to inform the development of the implementation experiences section of the the implementation report.

[15] HomVEE applies the term "developed-world context" to studies in countries that had high incomes in the year the manuscript was published, according to the World Bank Indicators list (World Bank 2020). For unpublished manuscripts, HomVEE will use the year the research was submitted to the call for research. However, research on **evidence-based** models conducted outside of the United States (but still in a developed-world context) is excluded unless HomVEE resources in a given year permit review of that research.

[16] If HomVEE cannot assess which model the manuscript is affiliated with based on its full text, and if no stakeholder indicated the model affiliation for the manuscript as part of the call for research, the team places the manuscript on hold until the author or developer chooses to identify the model to HomVEE in response to a subsequent call for research.

[17] Beginning with the 2021 review, HomVEE will implement a 20-year moving window for database searching. Two categories of older research remain eligible for HomVEE, as described earlier in this chapter: older research that HomVEE has already reviewed, and research submitted at any time (that is, since HomVEE's inception and moving forward) through the call for research. For models prioritized in 2018 and earlier, HomVEE also did a focused search reaching back to 1979. Because so few manuscripts published before 1979 related to models prioritized in recent years, starting with the 2019 review HomVEE limited the focused search to manuscripts reaching back to 1989 or later.

- The manuscript did not present findings from primary research. Primary research includes authors' own analyses of secondary data, but it does not include manuscripts that do not report original findings. Examples of the latter are literature reviews or meta-analyses.

The review uses a pair of databases (RefWorks and SharePoint) to catalog manuscripts and as a management tool to track the literature search, screening, and review process.

The screening process focuses on identifying entire manuscripts that are ineligible for review; however individual findings might be found ineligible for review during the review process even if a manuscript screens in based on other eligible findings. For example, findings that estimate indirect effects would be screened out during the review phase if the manuscript contained other findings that were eligible for review. For additional details, see Chapter III.

2. Calculate prioritization scores

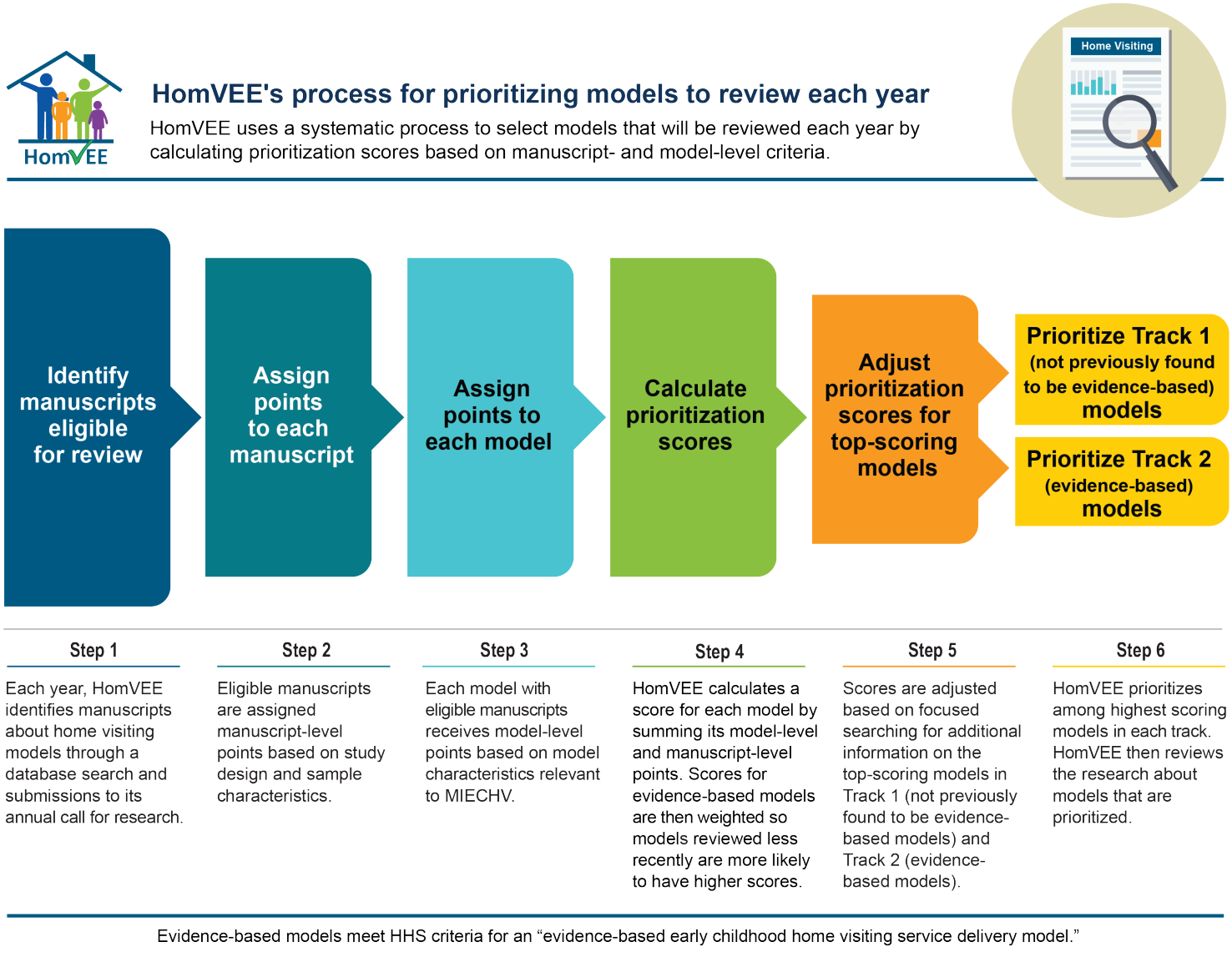
Given limited resources, HomVEE prioritizes eligible home visiting models to review. HomVEE prioritizes and reviews related versions (commonly referred to as adaptations) of a model together.

HomVEE's **prioritization process** (Exhibit II.4) reflects HomVEE's emphasis on reviewing well-designed impact studies, examining outcomes of interest to HHS, and aligning to MIECHV Program criteria. HomVEE aims to identify new evidence-based models while continuing to update reports on evidence-based models that HomVEE previously reviewed in order to ensure reported findings are up to date to the extent possible.

What HomVEE prioritizes each year depends on the available project resources, as well as on the prioritization score HomVEE calculates for each model using a combination of manuscript and model characteristics. Regardless of whether HomVEE reviews a model in a given year, the team will include the model and its associated manuscripts in the prioritization process in subsequent years, although no model will be reviewed in two consecutive years (see Select models for review, below). The MIECHV Program may coordinate with HomVEE to prioritize review of promising approaches implemented and evaluated under a MIECHV grant.[18]

[18] As per MIECHV statute, a home visiting service delivery model that qualifies as a promising approach (1) conforms to a "promising and new approach" to achieving specified benchmark areas and participant outcomes, (2) has been developed or identified by a national organization or institution of higher education, and (3) will be evaluated through a well-designed and rigorous process (see Social Security Act, Section 511 [42 U.S.C. 711]).

Exhibit II.4. HomVEE's prioritization process



Calculating prioritization scores involves four steps:

1. Apply manuscript-level criteria
2. Apply model-level criteria
3. Calculate prioritization scores
4. Adjust prioritization scores

Next, we describe these steps in the prioritization process and how the process differs for models in **Track 1** (models that HomVEE has not previously found to be evidence based) and **Track 2** (those that HomVEE has already reviewed, and that already are evidence based).

a. Step 1. Apply manuscript-level criteria

Manuscript-level criteria reflect HomVEE's emphasis on well-designed impact studies, outcomes of interest, and alignment with criteria in MIECHV's authorizing statute. A study's findings might be presented across several manuscripts, as discussed in Chapter I. HomVEE applies points to each manuscript that reports (1) previously unreviewed findings or (2) previously reviewed findings about well-designed, well-executed research on a model that HomVEE has not previously found to be evidence based.[19] HomVEE bases the points on information that authors provide in the title and abstract.[20] Specifically, when applying the manuscript-level criteria, HomVEE assigns points to each manuscript that is eligible for review based on the sample size and outcomes examined in the manuscript. At this step, HomVEE also considers aspects of the study that are described in the manuscript, including the study design, location of the sample, and population from which sample was drawn. Each model can earn up to 6.5 points for each impact study manuscript that is eligible for points (Exhibit II.5).

HomVEE assesses each manuscript separately and then sums the points for all manuscripts about a model. **Therefore, models with more eligible manuscripts tend to receive more points during this step.** This increases the prioritization scores for models with larger volumes of unreviewed research.

[19] HomVEE defines well-designed, well-executed research as manuscripts with at least one finding that rates moderate or high (see Chapter III), which suggests that some or all of the findings observed were due to the early childhood home visiting model and not to other factors. Other manuscripts about the same study could have different ratings, and even rate low. Manuscripts in which all findings rate low are not included in the point total for the model, even if other manuscripts about the same study have high or moderate ratings. HomVEE focuses on individual manuscripts, rather than studies because one study may span years or decades. Individual manuscripts reflect the volume of new research being produced about a model and the current state of the evidence base.

[20] At this stage in the prioritization process, HomVEE uses information provided in the title and abstract because it is not feasible to review the full texts of all manuscripts identified in a given year. As described in Step 4 of the following prioritization process, HomVEE rescreens manuscripts about top-scoring models using the full texts, and refines the manuscript-level criteria and points for those models accordingly.

Exhibit II.5. HomVEE manuscript-level prioritization criteria and associated points

Criterion	Points	Notes
Study design	2 to 3 per manuscript	3 points for each manuscript about a randomized controlled trial, single-case design, or regression discontinuity design (because these designs are eligible for HomVEE's highest rating). 2 points for each manuscript about a non-experimental comparison group design (because this design is eligible for HomVEE's moderate rating, at best).
Sample size	1 per manuscript	Total sample size reported in manuscript contains 250 or more pregnant women and/or families.
Outcomes of interest	1 per manuscript	Manuscript examines outcomes in one or more of the following domains for which HomVEE has seen comparatively less research over time: family economic self-sufficiency; linkages and referrals; reductions in child maltreatment; and reductions in juvenile delinquency, family violence, or crime.
Sample location	0.5 per manuscript	The entire sample reported in the manuscript lives in the United States.
Indigenous population	0.5 per manuscript	The entire sample reported in the manuscript is an indigenous population living in or outside the United States.
Priority population	0.5 per manuscript	The entire sample reported in the manuscript belongs to one or more priority populations named in the MIECHV authorizing statute.[a]

[a] According to Social Security Act, Section 511 [42 U.S.C. 711], 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

b. Step 2. Apply model-level criteria

Model-level criteria include factors that are related to eligibility requirements for the MIECHV Program. This increases the likelihood that models potentially eligible for MIECHV funding will be prioritized. HomVEE assigns model-level points to the model overall, based on information from model websites, information a model developer has supplied, and previous HomVEE reviews. The model receives one point if a criterion is true for that model (or, for any related version of the model). HomVEE may contact manuscript authors or model developers to confirm publicly available information. This process is identical for Tracks 1 and 2. Models can earn up to 4 points in this step, based on factors described in the MIECHV authorizing statute (Social Security Act, Section 511 [42 U.S.C. 711]):

Exhibit II.6. HomVEE criteria for model-level points

Criterion	Possible points	Notes
Associated with national organization or institution of higher education?	1	Organizations can be in or outside the United States.
Currently serving or available to serve families?	1	In assigning this point, HomVEE supplements information from developers with information from web searches and review of communication that developers and authors have submitted.
Implemented for at least three years?	1	Models can receive this point even they are not currently active. In assigning this point, HomVEE supplements information from developers with information from web searches, manuscripts, and review of communication that developers and authors have submitted.
Implementation support available in the United States?	1	HomVEE assumes international models support United States replication if that model has already been implemented in the United States, or if developers notify HomVEE that they would support United States implementation.

Note: HomVEE prioritizes and reviews related versions (commonly referred to as adaptations) of a model together. All related versions of a model receive one combined prioritization score. Each grouping of related models can receive a maximum of one point for each of the above criteria.

c. Step 3. Calculate prioritization scores

After assigning manuscript- and model-level points, HomVEE sums all points across both of these levels to calculate a model's point total and then applies a weight (that is, a multiplier) to the point total, as shown in Exhibit II.7.

Exhibit II.7. Weighting HomVEE prioritization scores to prioritize models for review

Model status	Weight	Rationale and notes
Reviewed the previous year	0	This ensures that no model is reviewed in two consecutive years.
Not reviewed the previous year, Track 1	1	The point total is equal to the final model prioritization score.
Not reviewed the previous year, Track 2	Weight = [1 + 0.1 * (current year – release date of prior report)]^2	Weight is based on the number of years since HomVEE last reviewed the model and released a report. For example, a model considered for review in 2020 that was last reviewed in 2016 would receive a weight of [1 + 0.1 * (2020 – 2016)]^2 = 1.96. A model considered for review in 2020 that was last reviewed in 2018 would receive a weight of [1 + 0.1 * (2020 – 2018)]^2 = 1.44. As this example illustrates, models reviewed less recently receive higher weights.

Note: HomVEE prioritizes and reviews related versions (commonly referred to as adaptations) of a model together. All related versions of a model receive one combined prioritization score. Each grouping of related models has its score weighted as shown above. In the first year of implementing these Version 2 procedures and standards, HomVEE will allow for models reviewed in 2020 to be considered for prioritization in 2021. Thereafter, HomVEE will apply the weights shown in this table.

After weights are calculated, a model's prioritization score is calculated as follows:

Prioritization score = (Manuscript-level points + Model-level points) * Weight

The weights ensure that no model is reviewed in two consecutive years. All Track 1 models are weighted equally, but Track 2 models are weighted so that models HomVEE reviewed less recently have a higher likelihood of being prioritized for review than models that have been reviewed more recently. This permits the review of evidence-based models to be updated periodically as new research on the model or one of its related versions emerges.

d. Step 4. Adjust prioritization scores

In the final step of calculating prioritization scores, HomVEE adjusts the scores based on a more thorough screening for research on top-scoring models. HomVEE sorts models from the highest to lowest score separately within each track. HomVEE then does two things to refine the point totals for the top-scoring models in each track. First, HomVEE conducts a second, focused database search on model names to identify additional manuscripts that were not identified using HomVEE's main keyword search terms. Second, after identifying additional manuscripts, HomVEE examines the full texts of all screened-in manuscripts and updates the number of points assigned to each manuscript based on information available from the full texts. The model's corresponding prioritization score is updated as well. This step updates scores to include information that was relevant to prioritization, yet was missing from manuscript titles and abstracts.

3. Select models for review

After calculating prioritization scores, HomVEE selects models for review. To do this, HomVEE re-sorts models separately in each track, using the adjusted prioritization scores. Then, HomVEE selects models from each track, starting with those with the highest scores and moving down the list in order of prioritization score. In any given year, the number of models prioritized for review depends on available project resources. The MIECHV Program may coordinate with HomVEE to prioritize review of promising approaches implemented and evaluated under a MIECHV grant.[21]

After selecting a model to review, HomVEE generally reviews all eligible, new manuscripts from impact studies about that model, including research on its related versions. However, HomVEE will not review research conducted outside the United States on a Track 2 model unless: (1) review resources for that year permit, or (2) the research was conducted with indigenous communities outside of the United States. This is because, when resources are limited, HomVEE aims to prioritize review of studies that are more likely to resemble the context in which MIECHV grantees might be implementing home visiting models. Research conducted outside the United States is less relevant than research conducted within its borders. However, research in indigenous communities is always of interest to HomVEE given the existence of a separate Tribal MIECHV program. If studies conducted outside the United States are not reviewed, the HomVEE website will clearly indicate which research was and was not included in the updated report.

[21] 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).

B. Conduct review

To be confident that home visiting models are effective, HomVEE needs to determine which research is well designed and executed. HomVEE does this in two steps, which are described in more detail in the sections below.

1. First, the review team asks: **Is the research well designed and executed? In other words, how confident can readers be that the findings were caused by the home visiting model and not by other factors?** As described in Section B.1, reviewers use a standard review protocol to evaluate the research design and methodology of eligible manuscripts about impact studies (see Chapter III for details on the standards). HomVEE reviews and rates the quality of an impact study described in a manuscript, according to each manuscript's **findings** (defined earlier in Exhibit I.4). The rating is an assessment of the strength of the research design behind the finding, which HomVEE characterizes as high, moderate, or low (Exhibit II.8). **Manuscripts receive an evidence rating based on the highest evidence rating of any one finding in the manuscript.** A high-rated manuscript may also have moderate- and even low-rated findings within it (for example, this could occur if the rate of attrition differs for different outcomes).
2. Then, HomVEE asks: **Based on well-designed, well-executed research only, was the home visiting model effective?** In this step, the team looks across all findings from well-designed research on a model to examine the direction and statistical significance of effects that authors find. Based on that, HomVEE determines whether the model is evidence based. Section B.2 describes these steps, and how stakeholders can request reconsideration of a model's evidence rating.

1. Review impact research

HomVEE follows a predefined process for reviewing and rating manuscripts, as described in Section B.1.a below. Occasionally, this process involves supplemental information that authors give HomVEE, or that HomVEE requests from authors, as defined in Section B.1.b.

a. Review and rating process

To ensure a review is as complete and accurate as possible, two certified reviewers review each manuscript. The first reviewer evaluates all of the eligible findings in the manuscript (see Chapter III, Section A of the handbook for information on how HomVEE determines which outcomes are eligible for review); rates them; assigns an overall rating to the manuscript of high, moderate, or low based on the highest rating of any of the study findings reported in the manuscript (Exhibit II.8); and records the results of the review. Then, a second reviewer, usually one more experienced with HomVEE or another systematic review with similar standards, examines the manuscript 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. Finally, the contractor's review team leader or deputy leader confirms all consensus rating decisions, consulting the HomVEE project leadership team as needed.

HomVEE reviewers examining a manuscript may use information learned from other manuscripts on the same study, as well as information provided by the author to assign an accurate rating, especially with respect to questions of compromised randomization.

Exhibit II.8. HomVEE's evidence ratings for findings and manuscripts

Rating	Interpretation
Rating findings	
High	There is strong evidence that at least one finding reported in the manuscript is attributable to the intervention that was examined.
Moderate	There is some evidence that at least one finding reported in the manuscript is attributable, at least partly, to the intervention that was examined. However, other factors not accounted for in the study might also have contributed to the finding.
Low	There is little evidence that the reported finding is attributable, partly or as a whole, to the intervention that was examined.
Rating manuscripts	
High	At least one finding in the manuscript is rated high according to HomVEE standards.
Moderate	At least one finding in the manuscript is rated moderate according to HomVEE standards (but no findings in the manuscript rate high).
Low	All findings that were eligible for review in the manuscript rate low.

Note: If multiple finding-level ratings apply to a given manuscript, HomVEE assigns the highest rating of any finding in the manuscript because of the strength of evidence that is attributable to at least one finding in the manuscript. HomVEE notes which findings rated lower than that and why in reporting the results of the review, and does not report the details of low-rated findings.

b. Incorporating information from authors and stakeholders

HomVEE reviewers often rely on clarifying information that authors and other stakeholders provide when reviewing a manuscript. Typically, this information comes in response to an author query that HomVEE initiates, as described in the next section. Sometimes, stakeholders give HomVEE other supplemental information. When that occurs, the timing and approach to using that information depend on when and for what purpose the stakeholder submitted it, as described in Section 1.b.ii.

i) Author queries

Some manuscripts are missing information that reviewers need to determine manuscript ratings, such as information on attrition or the baseline equivalence of the intervention and comparison groups (see Chapter III for definitions). In these cases, HomVEE sends queries to authors to request the missing information. Authors have one week to respond. HomVEE adjusts the rating of the relevant finding(s) and the manuscript based on authors' responses to these queries. If the authors do not respond, or do not provide the necessary information, HomVEE assigns a rating to the finding(s) and the manuscript based on the available information.

There are several situations in which HomVEE does not send queries to authors to request missing information:

- **The manuscript is missing information that HomVEE needs to determine a rating, and the manuscript provides no indication that the study collected the missing information.** For example, if reviewers need to assess whether race/ethnicity were equivalent in the intervention and comparison groups at baseline, and the study makes no mention of having collected this information, HomVEE will not query authors for this information.

- **The manuscript is only missing certain details about findings that already rate as moderate or high and that HomVEE would prefer to report, but HomVEE can skip reporting that information.** In these cases, HomVEE rates the findings based on available information, but may report that certain details (such as intervention and comparison group means, or effect sizes) are not available.
- **The query would require the author to perform new analyses. Generally, HomVEE only requests clarification of information that the author implies having, but did not explicitly include, such as a statement in the manuscript that groups are equivalent without corresponding test statistics.** Rarely, HomVEE requests that authors conduct new analyses for HomVEE review team. This only happens in two circumstances: (1) reviews of certain RCTs and non-experimental comparison group designs (NEDs) that use repeated measures analyses (Chapter III, Section 4.b in this handbook), and (2) reviews of structural equation models that are missing a diagram of the model (Chapter III, Section 4.c in this handbook).

ii) HomVEE's procedures for handling supplemental information

HomVEE accepts supplemental information only under specific circumstances. Supplemental information can take two forms: new information and new research. **New information** may discuss a study's methods or procedures. HomVEE incorporates that information only if (1) it is provided in direct response to an author query (see below) and (2) authors submit it in time for reviewers to examine it during the same annual review cycle in which HomVEE issued the query. Otherwise, authors must wait until HomVEE releases its annual review results for the model described in the manuscript in question. Then, they may follow the process for requesting a reconsideration of evidence to ask HomVEE to examine supplemental information about methods or procedures after the release of the annual review results.

New research could be additional findings or new analyses of research in a previously reviewed manuscript, or it could be an entirely new set of findings. HomVEE treats all new research as a submission to the following year's call for research, unless it consists of new analyses conducted at the explicit request of HomVEE.[22]

Exhibit II.9 specifies how HomVEE handles various situations involving supplemental information.

[22] HomVEE requests new analyses only in rare circumstances. For example, when reviewing repeated measures analyses, HomVEE may (rarely) ask authors to calculate adjusted time-point findings (if the manuscript does not already report time-point findings, and HomVEE reviewers are unable to calculate them from information in the manuscript). See Chapter III, Section C.3 for details. HomVEE also requests that authors of manuscripts about studies with single-case designs (SCDs) submit their raw data in graphical or tabular format, to support analyses that will calculate a design-comparable effect size (see Appendix D for details on standards and procedures for reviewing these studies).

Exhibit II.9. HomVEE's treatment of supplemental information provided by study authors and other stakeholders

Type of supplemental information	HomVEE treatment
New information OR new research that *directly answers* the questions HomVEE poses in an author query as reviewers examine a manuscript	HomVEE incorporates supplemental information in response to an author query into an active review (including any new analyses requested by HomVEE), if that information arrives by the deadline for the response to the author query. If information arrives after the deadline, HomVEE will incorporate it only if doing so is possible at that point in the annual review cycle. If it is not possible, HomVEE will finalize the review without considering the information provided by the author(s). HomVEE would examine the late-arriving answers only if authors appealed the manuscript rating after HomVEE published the review results.
New research: Updated version of an existing manuscript	HomVEE treats updated versions of existing manuscripts (for example, a published journal article based on an unpublished manuscript previously submitted to HomVEE) as a submission to the next call for research. If the call for research for the current review year is closed, HomVEE will not incorporate the updated version into the current review. As of the close of the call for research that follows the submission of the updated version, HomVEE incorporates the unpublished version into screening and prioritization decisions for that calendar year. When HHS next prioritizes the model that the manuscript examines, HomVEE compares the new version of the manuscript to the earlier one to look for (1) differences that would change the rating and (2) any findings that were not reported in the earlier version. HomVEE includes any updates from either of these two with the model report update in the fall of the year that the model is prioritized.
New research: Additional manuscript for consideration by HomVEE screeners and reviewers	HomVEE treats newly submitted manuscripts as a response to the next call for research. If the call for research for the current review year is closed, HomVEE will not incorporate the new research into the current review. As of the close of the call for research that follows the submission of the updated version, HomVEE incorporates the new manuscript into screening and prioritization decisions for that calendar year. When HHS next prioritizes the model that the manuscript examines, HomVEE will screen and review the manuscript and include the results in HomVEE products released that fall. This rule applies regardless of whether the additional manuscript is a stand-alone submission, is packaged with a response to an author query that HomVEE initiated, or is packaged as part of a request for re-review that a stakeholder submits after HomVEE publishes the review results.
New research that the author volunteers (not in response to HomVEE query)	Authors sometimes submit new research in the form of additional findings or new analyses of research in a previously reviewed manuscript. Authors also occasionally re-analyze existing outcomes in new ways (such as transforming the data or changing the covariates). Stakeholders submitting additional findings must clearly indicate to HomVEE that they are submitting new analyses or findings. HomVEE treats additional findings and new analyses as a submission to the call for research. As of the close of the call for research that follows submission of this additional information, HomVEE incorporates that information, as a new manuscript, into the annual screening and prioritization process. HomVEE will examine the new manuscript or additional analyses when HHS next prioritizes the model that the manuscript examines.

Type of supplemental information	HomVEE treatment
New information: Additional details the author or another stakeholder offers about a study's methods and procedures (not in response to HomVEE query)	New details on methods and procedures may include: attrition information, data on the baseline equivalence, and statistics (such as *p*- values or effect sizes) reflecting the authors' analysis. If the author or stakeholder provides additional details as the basis for an appeal about the manuscript's published rating after HomVEE publishes review results, HomVEE will generally re-examine the manuscript and report the team's findings to the stakeholder within 60 days (see Chapter II, Section B.2.d on requests for reconsideration). For example, the additional details could support the author's case that HomVEE standards were erroneously applied, could be information that the author does not think HomVEE considered, or could be information that the author provided after an author query deadline. HomVEE will not examine new details about study methods and procedures for a manuscript HomVEE has already reviewed unless they arrive through the appeal process. Examples of communications that this exclusion applies to are: a cover message accompanying a submission to the HomVEE call for research, an informational message the author opts to share with HomVEE, and supplementary details about the manuscript or study that the author offers when responding to a HomVEE author query.

2. Assess model effectiveness

An assessment of model effectiveness relies on all available research about the model so that HomVEE users can be confident that the review is comprehensive. After reviewers have rated all of the identified manuscripts for a prioritized model, HomVEE synthesizes the high- and moderate-rated findings. First, reviewers assess the direction and statistical significance of each finding. Next, HomVEE synthesizes the evidence separately for each of the eight HomVEE outcome domains. Then, HomVEE examines findings across manuscripts before making an overall rating of model effectiveness. The sections below describe each of these steps, and also explain how stakeholders may request reconsideration of model ratings.

a. Identify direction and statistical significance of findings

First, HomVEE assigns one of three categories to each high- or moderate-rated finding:

- **Favorable.** A finding showing a statistically significant impact on an outcome measure in a direction that is beneficial for children and parents. An impact could be statistically positive or negative, and is determined to be "favorable" based on the result. For example, a favorable impact could be an increase in children's vocabulary or in daily reading to children by parents, or a reduction in child maltreatment or maternal depression.
- **No effect.** Findings are not statistically significant.
- **Unfavorable or ambiguous.** A finding showing 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 is determined "unfavorable or ambiguous" based on the result. Although some outcomes are clearly unfavorable, for other outcomes it is less clear which direction is desirable. For example, an increase in children's behavior problems is clearly unfavorable, whereas an increase in the number of days mothers are hospitalized after birth is ambiguous. It could be viewed as unfavorable because it indicates that mothers have more health problems, but it could also indicate that mothers have increased access to needed health care because they are participating in a home visiting program.

HomVEE considers a finding to be a **statistically significant impact** if the *p*-value of a two-sided hypothesis test of whether an effect is equal to zero is less than 0.05. That is, the finding is likely to be due to the home visiting model, rather than due to chance.

b. Determine whether study samples overlap

Before assigning model effectiveness ratings, HomVEE considers how manuscripts group into studies, and whether the study samples overlap.

i) Considering how manuscripts group into studies and sample overlap for most manuscripts

Two situations may indicate that manuscripts are part of the same study:

- **Identical samples:** Two or more manuscripts that report results from an analytic sample whose entire group of participants consists of the same sample members. For example, this could be two manuscripts on the same intervention and comparison group that report findings on different outcomes.
- **Overlapping samples:** Two or more manuscripts in which the intervention groups or comparison groups have at least some sample members in common. For example, researchers following participants over time who lose some participants in follow-ups would have a sample for the later follow-up that overlaps with the sample from the earlier follow-up.

In contrast, manuscripts may report findings from **distinct (or non-overlapping) samples,** in which there are no sample members in common. This situation means that the manuscripts are about distinct studies (for example, one study conducted from 2002 through 2004 in state A and another conducted with a newly recruited sample from 2008 through 2010 in state B). When applying the HHS criteria to identify evidence-based models, HomVEE carefully examines whether the manuscripts it has reviewed have identical, overlapping, or distinct samples. Reviewers also may use information from one manuscript about a study to contribute to decisions about the rating of another (for example, if one manuscript describes that randomization was compromised and another, about the same sample, does not include that detail.)

ii) HomVEE procedures for subgroup research

At this point, HomVEE also considers whether the findings come from subgroup analyses, based on a careful examination of subgroup research and the studies the subgroups come from. Subgroup research is important for HomVEE because a model can earn an evidence-based rating through findings from subgroups (Exhibit II.11). Therefore, HomVEE exercises care in identifying subgroup research and understanding how the subgroup relates to the overall study sample.

HomVEE defines a subgroup as a subset of the *overall sample examined in a study*—that is, an **analytic subgroup** (see Exhibit I.4).[23] Notably, this is different from defining subgroup as a subset of the *overall population.* Although researchers may examine an analytic subgroup in hopes of making inferences about a subset of the population, the goal of the HHS criteria is to ensure that program impacts are replicated consistently for an outcome domain. Such replication is what gives HomVEE confidence that evidence of effectiveness is not due simply to chance. Thus, if a model is evidence based due to subgroup findings,

[23] For research reviewed by HomVEE, subgroups may be defined a priori (that is, before the research begins) or post hoc (after the research is underway). HomVEE applies the same standards for assessing the quality of research and the same rules about replicability to subgroups regardless of when researchers define the subgroup.

this means that research in which that subgroup was similarly defined in relation to the broader sample had consistent, favorable (statistically significant) findings in distinct study samples.

Subgroup results may be nested within a manuscript (for example, results from teenage mothers when the overall results in the manuscript are from mothers with a range of ages), or they may be the main focus of a manuscript (for example, a manuscript focusing on results from teenage mothers when the overall study sample included mothers with a range of ages). HomVEE treats both of those as analytic subgroup analyses. HomVEE's definition means that not all analyses restricted to a certain characteristic are subgroup analyses. For example, results from teenage mothers are not an analytic subgroup analysis when the overall study only enrolled teenage mothers, even though teenage mothers are a subgroup of the population of mothers as a whole.

Because HomVEE's mission is to identify evidence-based models, and to use project resources judiciously, the project only ***reviews*** research on **replicable subgroups** (if it meets other eligibility criteria defined in Chapter II, Section A of this handbook), and HomVEE typically only ***reports*** review results for **replicated subgroups** (see Exhibit II.10). HomVEE uses the same set of standards to rate the quality of research for replicated subgroup findings. For example, if the subgroup replication straddled the timing of an update to HomVEE standards, HomVEE reviews both sets of subgroup findings using the newest standards (even if other, full-group findings from the older manuscript were reviewed using older HomVEE standards).

Few subgroups have been replicated to date in research on evidence-based models, but HomVEE will report findings from replicated subgroup analyses when they appear.[24] In addition, beginning with research examined in the 2021 annual review, HomVEE will list nonreplicated subgroups that researchers have examined in each manuscript (without rating the quality or reporting the specific details of those findings).

[24] The challenge of reproducibility of impact findings is an important consideration in the movement for open science. Although HomVEE's focus on replicated subgroups addresses this for subgroup findings, HHS' overall criteria for effectiveness (see Exhibit II.11) also reward reproducibility by requiring that, if a single study finds only one favorable finding for a model, a second study must also have a favorable finding in that same domain.

Exhibit II.10. HomVEE procedures for reviewing and reporting subgroup research

HomVEE reviews replicable subgroups, and reports subgroup results only once the results are replicated. For HomVEE, those terms are defined as follows:

- **Replicable subgroups** are defined by a characteristic that a different study could replicate with a non-overlapping sample. Most subgroups are replicable in theory. However, HomVEE does not consider subgroups defined by cohort or time (for example, a subgroup of mothers enrolled in 1995 in a study that included mothers enrolled across several years) to be replicable in subsequent studies, and therefore does not review time-based subgroups. Similarly, HomVEE will only consider a subgroup defined by location to be replicable if the location was selected based on defined characteristics (for example, county with the highest teen birth rate in the state in a study conducted in several counties). Location-based subgroups defined by a location name (for example, Adams County in a study conducted in several counties) will not be reviewed because the HomVEE team cannot confidently verify whether the subgroup sample in a subsequent study in that county overlaps with the first study when the team applies HHS criteria.
- HomVEE will report subgroup results only from a **replicated subgroup,** one that has an identical definition in two non-overlapping research samples. For example, a study examining a subgroup of primiparous teenagers is not replicated by a study examining primiparous women of all ages. This approach is consistent with the HHS criteria's emphasis on observing effects across independent samples.

c. Determine the model's effectiveness, according to HHS criteria

Finally, based on the direction and adjusted statistical significance of the findings in each domain, HomVEE assesses whether each model meets the HHS criteria for an "evidence-based early childhood service delivery model" (Exhibit II.11).[25] Although HomVEE prioritizes and reviews related versions of a model together, each version of a model receives its own assessment of evidence of effectiveness. A model may be evidence based on the strength of subgroup findings alone *only if* the research about it satisfies all of the subgroup criteria. To operationalize the HHS criteria related to studies, and because study findings may be reported across several manuscripts, HomVEE rates manuscripts based on the highest rated finding reported in that manuscript. Any high- or moderate-rated finding from a study about a model is considered as part of the evidence base for that model. Notably, for models with research solely from either RCT or SCD studies, additional criteria apply (see Exhibit II.11). The additional criteria for RCTs to be from peer-reviewed journal articles and to have sustained findings, align with MIECHV statutory requirements. The HHS criteria for an evidence-based model have no additional requirements for RDD or NED studies.

[25] Periodically, HomVEE also conducts a special review focused on home visiting research with tribal populations. This focuses both on manuscripts about models that were used in tribal communities and manuscripts that identified 10 percent or more of the sample as tribal; the review of research with tribal populations is distinct from the HomVEE annual review that incorporates research across populations. The HomVEE review of research with tribal populations aims to identify evidence-based models based on research from either (1) a sample composed entirely of tribal participants or (2) impacts reported by ethnicity/tribal community affiliation, with those subgroup findings replicated in another distinct sample. It also reports on implementation of early childhood home visiting models in tribal communities.

Exhibit II.11. HHS' criteria for an "evidence-based early childhood home visiting service delivery model"

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-rated 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-rated 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 for the study or (2) if found for subgroups but not for the full sample for the study, be replicated in the same domain in two or more studies using non-overlapping analytic study samples. Additionally, following the MIECHV-authorizing statute, if the model meets the above criteria based on findings from randomized controlled trials only, then two additional requirements apply. First, one or more favorable (statistically significant) impacts must be sustained for at least one year after program enrollment. Second, one or more favorable (statistically significant) impacts must be reported in a peer-reviewed journal.

For results from single-case design (SCD) research, different criteria apply.

Note: HomVEE intends to adopt WWC Version 4.1 standards for reviewing SCD research. Therefore, the SCD criteria from HomVEE's original procedures and standards are no longer applicable. Revised criteria for SCD research and opportunities for public comment will be forthcoming.

Note: HomVEE allows the two requirements about sustained and peer-reviewed impacts listed after the bullets to be satisfied by findings from different studies, provided the quality of these findings is rated as moderate or high. These criteria are consistent with the MIECHV statutory requirements: Section 511 (d)(3)(A)(I)(I).

d. Requests for reconsideration of model evidence determinations

Once HomVEE publishes the results of its review, if a state/territory/tribal program administrator, researcher, model developer, or other interested individual believes that (1) in applying the criteria to determine whether a particular model is evidence based, HomVEE made one or more errors; and (2) if these errors were addressed, the model would be evidence based, the stakeholder may appeal HomVEE's rating of a manuscript according to HomVEE standards (see Chapter III of this handbook) or application of the HHS criteria to a model. Stakeholders with these concerns should submit their inquiry to HomVEE@acf.hhs.gov. Inquiries are accepted only through this e-mail address. Individuals may request reconsideration of the evidence-based determination based on misapplication of the HHS criteria, missing information, or errors on the HomVEE website. Also, once HomVEE publishes its rating of a manuscript, authors may appeal to have HomVEE revisit the rating to incorporate specific information from late responses to an author query (see Exhibit II.9).

HHS will consider these requests as they arrive. If HHS approves HomVEE to investigate the request, to ensure independence from the original review, a re-review team composed of members external to the original contractor review team's organization will conduct a new, independent review of the manuscript(s) in question under the HomVEE standards that were in place when the manuscript was reviewed the first time. The re-review team will provide assurance that they do not have any actual or perceived conflicts of interest. This re-review team will not include members who were involved in the original review. As with the original review, the re-review team members will be certified and trained in the HomVEE standards. The re-review team will use the original empirical articles (see the model reports), any new *information* (but not new *research*) submitted as part of the request by the individual

raising the concern, and the original contractor review team's reports, and will make any needed queries to the original contractor review team (see Exhibit II.9).

HomVEE aims to issue a final decision as to whether the standards were accurately applied within 60 days of the submission of the request for review. Following the final decision, the requester will be notified of the decision, and, if necessary, HomVEE will make any necessary adjustments to the model effectiveness research reports or HomVEE website.

C. Report results

For each prioritized model, HomVEE produces reports that summarize HomVEE's assessment of the model's impact research and summarize the model's implementation.

1. Model effectiveness research reports

As stated, HomVEE's primary function is to help stakeholders understand which home visiting models are effective. After reviewing manuscripts and determining models' evidence of effectiveness, HomVEE summarizes this information in an effectiveness research report about the model on the HomVEE website. Each effectiveness research report indicates whether the model is an evidence-based model Then the report describes the model, impact study manuscripts and their HomVEE ratings, and presents summaries of findings from moderate- and high-rated manuscripts. For manuscripts that had at least one high- or moderate-rated finding, the reports also describe the study participants, the setting, a summary of the intervention and comparison group services, the characteristics and training of home visiting staff, and a listing of any subgroups examined (HomVEE will report findings from replicated subgroup analyses when they appear, see Chapter II, Section B). To emphasize the importance of open science practices, HomVEE also reports the study's funding source; author affiliation(s), including whether the author is the model developer; and whether the study was preregistered at ClinicalTrials.gov, the American Economic Associations registry for RCTs (soocialscienceregistry.org), or the Registry of Efficacy and Effectiveness Studies (https://sreereg.icpsr.umich.edu/sreereg/).

HomVEE's impact research reports also include additional details for individual findings that rate as moderate or high, including the following:

- Outcome name
- The direction and statistical significance of each effect
- Timing of the outcome measurement
- Sample size and description
- Means for the intervention and comparison groups and the difference between the two
- Effect size (a standardized measure of magnitude) [26]

[26] HomVEE does not require, but strongly encourages, reporting of effect size. HomVEE accepts measures of effect size provided by authors; if authors do not calculate effect size, HomVEE will do so if enough details are available. Then, HomVEE calculates effect size according to the approach defined by the What Works Clearinghouse (WWC) Procedures and Standards Handbook, Version 4.1. For RCT and NED studies, WWC bases their significance tests on an estimate of standard error that depends on sample size. For findings from continuous measures, which have a continuous set of potential values between the lowest and highest possible scores, HomVEE calculates effect sizes as Hedges' g. This is the ratio between the estimated impact of the intervention (the difference between the intervention and comparison group scores) and the standard deviation (the variation in scores) pooled across the

- How the outcome was measured
- How data were collected
- Reliability and validity properties (see Chapter III, Section B.4 of this handbook for HomVEE's face validity and reliability requirements)

If a manuscript includes high- or moderate-rated findings from a replicated subgroup (see Exhibit II.10), HomVEE also reports subgroup results.

2. Model implemementation summary reports

To provide more additional context for HomVEE's findings and to better inform stakeholders, HomVEE provides some information about model implementation on its website. For all prioritized models, HomVEE conducts Internet searches to find implementation materials and guidance available from home visiting model developers and national model offices. The team may also collect information about model implementation from effectiveness and implementation research identified through the literature search and screening process.

Model implementation summary reports mainly consist of detailed **implementation profiles** for each prioritized model includes an overview of the model (including any related versions of the model) and information about prerequisites for implementation, materials and forms, estimated costs, and model contact information. Model developers or national model offices are invited to review and comment on the profiles before their release.[27]

Implementation profiles are updated according to the model's evidence base and prioritization score:

- **Evidence-based models.** HomVEE may update the profile periodically (including years when the model is not prioritized for an updated effectiveness review) if the model (or any of its related versions) is evidence based.
- **Other models.** HomVEE updates implementation profiles when that model is next prioritized to have its effectiveness review updated.

Therefore, implementation profiles describe the latest information on how a developer thinks the model should be implemented; this may differ from the way it was tested.

intervention and comparison groups. For findings from dichotomous variables, which have only two possible values, HomVEE calculates effect sizes using the Cox index, which calculates an effect size for proportions. To avoid bias due to small sample sizes, WWC applies a sample size correction to effect sizes. See the WWC Procedures Handbook, Version 4.1 (U.S. Department of Education 2020a), Chapter VI for more details. For SCD studies, WWC calculates a design-comparable effect size (see Appendix D of this HomVEE Version 2 Handbook for more details).

[27] Prior to the 2021 review, for any evidence-based model, HomVEE also extracted and reported information about implementation experiences from the implementation studies and well-designed impact studies that are reviewed, as a single webpage in the implementation report. This information included the characteristics of participating families, location and setting, staffing and supervision, model components, model adaptations or enhancements, dosage, and lessons learned.

III. Standards for Rating the Quality of Impact Research

In this chapter, we describe the standards governing HomVEE's level of confidence that a given home visiting model caused the impacts observed in the research pertaining to it. Section A describes the standards HomVEE uses to confirm that the design and the findings examined in a study's manuscript are eligible for review. Assessing whether a manuscript uses an eligible design generally occurs at the screening stage, as discussed in Chapter II. Given that the design is eligible, assessing whether a manuscript has analyses and outcomes that are eligible generally occurs at the reviewing stage, also described in Chapter II. In contrast to the procedural focus of Chapter II, Chapter III Section A focuses on the technical details of research design that the review team considers during these screening steps. Then, Section B presents (1) the standards used to assess the rigor of the research design and (2) the requirements that reported findings must meet to receive a rating of high, moderate, or low. As described in Chapter II, the contractor review team assigns a rating (high, moderate, or low) to each finding within a manuscript that is eligible for review. On this basis, HomVEE determines the manuscript's rating. Specifically, the manuscript's rating equals the highest rating of any eligible finding in it.

Rating standards vary depending on a study's design. Occasionally, HomVEE reviews research with designs and analytic approaches that are comparatively less common in the home visiting literature. In Section C of this chapter, we describe HomVEE's approach to these designs and analytic approaches, which include repeated measures research and structural equation models. Finally, Section D describes HomVEE's approach to imputation and handling of missing data.

A. HomVEE's approach to determining which study designs, analyses, and outcomes are eligible for review

Reviewers examine each eligible manuscript on a study about a model that HomVEE prioritizes for review; they first assess whether the study reported in the manuscript uses an eligible design and analysis to test effectiveness and whether that analysis contains at least one eligible finding.

1. Eligible designs

HomVEE reviews research designed to clearly establish whether a home visiting model affects the outcomes of children and families. Designs eligible for review by HomVEE include randomized controlled trials (RCTs), and three types quasi-experimental designs, or QEDs: (1) regression discontinuity designs, (2) single-case designs, and (3) non-experimental comparison group (NEDs). NEDs compare an intervention group receiving a home visiting intervention to a comparison group that does not receive it, but in these designs assignment to the intervention and comparison condition does not happen randomly. Instead, it is the result of criteria determined by researchers. For example, researchers can create an intervention and a comparison group by statistically matching the comparison and intervention group members.

Impact research eligible for review by HomVEE includes

- Randomized controlled trials (RCT)
- Quasi-experimental designs (QED)
  - Non-experimental group designs (NEDs)
  - Single-case designs (SCDs)
  - Regression discontinuity designs (RDDs)

All of these designs can be subject to bias, but the risk of bias and how this bias affects the level of confidence one can have in the research findings is different for each design. Because of that, HomVEE takes into account the level of bias when assigning ratings to the research it reviews. For example, RCTs with no threats to the original assignment of sample members to intervention and comparison groups face the least amount of bias, and are therefore eligible for a high rating. NEDs, on the other hand, face a higher risk of bias than RCTs do. That is because NEDs assign sample members to intervention and comparison conditions through a process that is not random assignment. Consequently, this design cannot rule out the possibility that there are still some differences between the two conditions. Therefore, the highest possible rating an NED is eligible for is moderate. Next, we describe in detail the features of each design and the risk of bias it faces.

a. Randomized controlled trials

As the name suggests, **RCTs** assign sample members to intervention groups and comparison groups at random. Sample members can be assigned as individuals or as groups, depending on the study. For example, studies can assign sample members by household (as individuals or family units) or by county or ZIP code (as groups). When study participants are assigned as groups, the study design is called a **cluster RCT**. Each study participant must have at least some probability of being assigned to each study group; however, the probability of being assigned to each group does not need to be the same. HomVEE does not consider assignment of family units to be a cluster design.

Random assignment creates groups that are expected to be equivalent across both measured and unmeasured characteristics. These studies can provide strong evidence that differences in the outcomes of the intervention and comparison groups at the end of the study can be attributed to the intervention and not to preexisting differences between the groups (Shadish et al. 2002). This means that the groups are expected to be different from each other only in the sense that one group receives services from the home visiting model, and that those services would therefore explain any differences in outcomes between the groups. The main risk of bias for RCTs stems from sample members leaving the study (attrition), which might cause underlying differences in the sample members who remain in the study. That is, sample members who remain might be different from the ones who left because of qualities that contributed to them leaving.

Compromised randomization also causes a threat of bias in RCTs, and when this arises HomVEE reviews the research using standards for non-experimental comparison group designs. An RCT is compromised when either (1) some of the participants in the analytic sample are not part of the sample that was randomized (that is, when there are nonrandom additions to the sample); or (2) after random assignment, some participants are moved from the group to which they were originally assigned to another group (that is, there is reassignment). The presence of noncompliers does not compromise an RCT if the analytic sample includes participants in the groups to which they were originally assigned.[28] **Noncompliers** are study participants who receive services they were not supposed to receive (for example, participants who were randomly assigned to the comparison group receiving the services from the home visiting model offered to the intervention group).

[28] If a manuscript does not allow reviewers to clearly assess whether randomization was compromised, HomVEE will issue an author query to ask for the necessary information.

b. Non-experimental comparison group designs

Non-experimental comparison group designs (NEDs) use a nonrandom process to assign sample members to an intervention group and a comparison group. Sample members can be assigned through statistical techniques that are designed to match sample members in each group, so each group has similar measurable characteristics on average; or they can be assigned based on convenience, by assigning people to groups because they are nearby or available or otherwise convenient to include.

Unlike RCTs, in NEDs the study participants assigned to the intervention and comparison groups may differ on measured or unmeasured characteristics. For example, study participants assigned to the intervention group may have higher assessment scores before the home visiting starts than study participants assigned to the comparison group do. Thus, any differences in assessment scores at the end of the home visiting period cannot necessarily be attributed to the services. The main threat of bias in NED studies is that study groups may not appear similar, and therefore any observed effects of the intervention might be attributable to the differences between study groups and not to the intervention.

c. Regression discontinuity designs

In **regression discontinuity designs (RDDs)**, study participants are assigned to the intervention and comparison groups using a criterion as a cutoff point, and researchers compare participants who are some set distance above and below the cutoff point. For example, study participants may be assigned to the intervention and comparison groups based on whether they have an assessment score above or below a cutoff value. With an RDD, the effect of the intervention is estimated as the difference between mean outcomes of the intervention group members and comparison group members *at the cutoff point*, adjusting for the relationship between the outcomes and the variable used to create the cutoff (that is, the variable used to assign sample members to the intervention). The effect of the intervention as estimated with RDD methodology is considered to be unbiased if the methodology meets specific standards.

These standards include:

- The relationship between the outcome and the variable used to create the cutoff must be smooth (that is, continuous) at the cutoff point.
- The variable used to create the cutoff must not have been manipulated to influence assignment to the intervention group.

Appendix C describes in detail HomVEE's standards for reviewing RDDs, which have been adopted from Version 4.1 of the WWC Standards (U.S. Department of Education 2020b). The What Works Clearinghouse (WWC) is a systematic review of education research established by the Institute of Education Sciences in the U.S. Department of Education.

RDDs can result in intervention and comparison groups that are different from each other on both measured and unmeasured characteristics, even with an assignment process that is based on a cutoff. The main concern about bias in these designs is whether sample members appear similar at and around the cutoff so that differences between them can confidently be attributed to the home visiting intervention. (See Appendix C for details on how HomVEE assesses the risk of bias in impact estimates from RDDs.)

d. Single-case designs

Instead of assigning participants to intervention or comparison conditions (as done in the study designs described above), single-case design (SCD) home visiting evaluations assign the intervention and comparison conditions to a single family or a small group of families **during certain time periods**. In single-case designs, researchers follow each study family or small group of families across several points in time. For example, a study may examine three families who alternate multiple times between receiving and not receiving a home visiting model. The study then compares outcomes during the time periods when the study participants received the home visiting model to outcomes during the time periods when they did not. By using each individual or small group of individuals as their own comparison group, single-case designs ensure that the intervention and comparison groups have the same measured and unmeasured characteristics. SCDs are subject to potential bias from differences between time periods that are not part of the model. (Notably, the standards HomVEE applies are intended to identify well-designed studies that limit this risk. See Appendix D for the SCD standards and their associated reporting procedures, adopted from WWC Version 4.1).[29]

2. Contrasts that HomVEE reviews

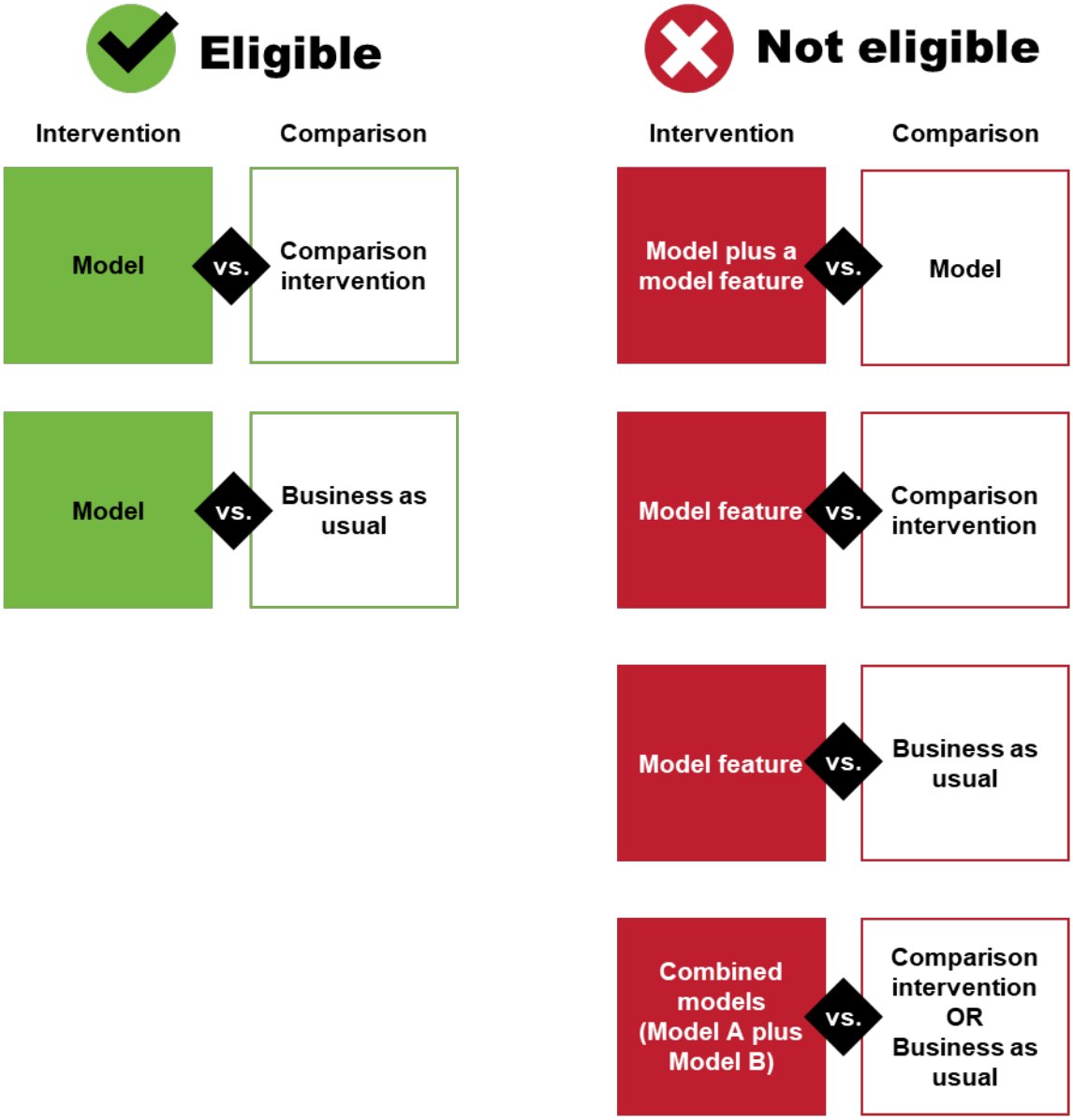
Precision home visiting research is valuable for answering questions about for whom and under what circumstances models work best. This work focuses on the components of home visiting services rather than on complex models of home visiting that are administered uniformly.[30] However, those questions are not the primary focus of HomVEE's annual review. HomVEE focuses on the primary research question of whether a given model is effective. Therefore, HomVEE excludes research that isolates the impact of models' features because that research does not answer the main question of whether an early childhood home visiting model is effective.

Specifically, research evaluating the impact of an early childhood home visiting model generally is eligible for review. This includes, but may not be limited to, research that compares an early childhood home visiting model to either a comparison intervention, or to business as usual, as depicted in Exhibit III.1. For HomVEE, business as usual means the typical services that routinely are available to the population under study, and are not an early childhood home visiting model.

[29] Although the previous standards instructed reviewers to use visual analysis of changes in the outcome over time and across conditions to characterize the findings from an SCD, the new standards calculate and use a design-comparable effect size to characterize the findings. Reviewers will still use visual analysis to assess whether an SCD study is well-designed. To calculate a design-comparable effect size as described in Appendix D, HomVEE will use data presented in the study if possible, or (only if necessary) contact the study authors to request raw study data.

[30] For more information on precision home visiting research, see for example: https://www.hvresearch.org/precision-home-visiting/

Exhibit III.1. Eligible and ineligible comparisons



Note: Eligible research about a model generally includes research about a version of the model. HomVEE defines *business as usual* as a condition characterized by typical services routinely available to the population under study.

Research evaluating the impact of an isolated feature or group of features is generally ineligible for inclusion in HomVEE's annual review. HomVEE focuses its resources on reviewing manuscripts about impact studies that answer the review's core question of whether an early childhood home visiting model is effective. Specifically, because knowing that a certain feature of a model is effective in isolation does not establish that a model consisting of multiple features (including the effective one) is effective overall. As such, research on single features should not contribute to the evidence base for an entire model. However, research reviewed by HomVEE might not be limited to the contrasts summarized in Exhibit III.1. For example, if a feature or group of features of a model also satisfies the definition of an early

childhood home visiting model, a study isolating the impact of that feature could potentially be treated as evidence for a separate model.

3. Ineligible and preferred analyses

HomVEE's mission is to determine whether research shows that a home visiting model improves outcomes for children and families. For this reason, certain types of analyses designed to answer questions other than whether a model is effective are not eligible for review, even if they are part of otherwise eligible impact analyses. Once HomVEE has confirmed the study uses an eligible design, a reviewer examines whether any analysis in the manuscript uses an eligible analytic approach. Two types of approaches are broadly ineligible for review by HomVEE: most mediating and moderating analyses, and all analyses that control for endogenous characteristics. (As discussed in Chapter III, Section C, specific types of cluster, repeated measures, and structural equation model analyses may also be ineligible for review.)

In addition, analyses of how the home visiting model affected *only* sample members who *received* it are de-prioritized if other, more preferred analyses are also reported in the same manuscript, as described at the end of this section.

a. Most mediating and moderating analyses

HomVEE focuses on research that answers the question: Is the home visiting model effective? Questions about the mechanisms behind how a model works, the settings where it might work best, and the populations who benefit the most from the intervention are outside of the scope of the HomVEE review. Although answers to these questions are important for understanding and improving home visiting models, the primary aim of the HomVEE review is to identify currently available models that are effective. Therefore, HomVEE mainly excludes two other types of analyses that are designed to answer questions that are slightly different from questions about model effectiveness:

- **Mediating analyses**, which investigate the process by which the home visiting model achieves its effects. These answer the question: How (or under what circumstances) does the model work? Researchers test these questions by conducting a path analysis, estimating certain types of multiple linear regression models, or running a structural equation model. Some mediating analyses that authors depict as structural equation models (see Section C.3 in this chapter for HomVEE's standards for reviewing SEMs) are eligible for review by HomVEE. All other mediating analyses are excluded by HomVEE.
- **Moderating analyses**, which investigate the ways that specific variables influence the effectiveness of the home visiting model. These answer the question, Does the model work equally well for different groups?[31] Researchers test these questions by including an interaction term in their model. HomVEE excludes analyses with continuous moderator variables. If a moderator variable is binary (for example, primiparous versus multiparous mothers), is not endogenous (see next section), and if authors provide enough data to complete the review, HomVEE may review moderating analyses as subgroup analyses. (See Exhibit II.10 for details on how HomVEE describes a subgroup.)

Other HomVEE products besides the annual review may report on mediating or moderating analyses, and the annual review that is the focus of this handbook is the main but not the only product of HomVEE. For

[31] Definitions derived from MacKinnon (2011).

example, HomVEE's *Evidence Says* brief on intimate partner violence included results from a manuscript that used a regression model to estimate the moderating effect of the presence of domestic violence on the effectiveness of home visiting interventions aimed at reducing child abuse and neglect (Eckenrode 2000).

b. Analyses that control for endogenous characteristics

Endogenous characteristics are characteristics of study participants that are defined by behavior that emerges after they learn whether they will be in the intervention group or the comparison group, or could theoretically be affected by a home visiting model. Therefore, there is a relationship between the assignment to the intervention and the endogenous characteristics (that is, they are not independent of each other). Consequently, analyses that control for endogenous characteristics can produce biased estimates of an intervention's effectiveness. For this reason, **HomVEE excludes analyses that control for endogenous characteristics.**

For example, consider a study measuring the effect of home visiting services on children's language and literacy skills. The study analysis statistically controls for a variable measured halfway through the home visiting service period that captures parent–child engagement—a characteristic that home visiting services could have impacted favorably. Therefore, the measure of parent–child engagement is capturing some of the effect of the home visiting model, and the remaining effect of receiving the model is biased (in this example, by understating how much home visiting can affect children's language and literacy). However, such questions are outside the scope of HomVEE's focus, and these analyses are therefore ineligible for review. In the same example, an analysis that considers the effects of home visiting services on parent–child engagement (that is, parent–child engagement is the outcome) and that does not control for any other endogenous characteristics is eligible for review.

Analyses of endogenous subgroups are also ineligible for review. These analyses use endogenous variables as a control in a specific way: by creating subsets (that is, subgroups) of the analysis sample defined by an endogenous binary or categorical variable and producing separate impact estimates for each subset. For example, consider an evaluation of the effect on infant death rates of a home visiting intervention that began during pregnancy, and whether the effect varies according to whether the infant was born preterm. Infant death itself is a valid outcome measure for the study. However, home visiting itself could have affected both preterm birth and infant death, so the analysis that examines the effect of home visiting delivered during pregnancy on infant death *for infants born preterm* does not isolate the effect of the home visiting model on infant death specifically. As a result, even with an experimental design, the intervention and comparison groups within subgroups defined by an endogenous variable lack equivalence and the estimated impact for the each subgroup captures only part of the effect of the intervention, leading to biased estimates of an intervention's impact for these groups (Colman 2012).

c. Treatment on the treated: Effect of home visiting on participating families

Authors of home visiting research may examine the effect of the **intent to treat** (ITT, or the effect of being offered the home visiting intervention) or the effect of the **treatment on the treated** (TOT, or the effect of actually receiving the home visiting intervention). Typically, manuscripts that HomVEE reviews examine the ITT estimate. When a manuscript reports both the ITT and TOT estimates, HomVEE focuses its review on the ITT estimate, because those estimates more realistically depict the average magnitude of

the effect that a program replicating the model would observe.[32] If authors report only TOT estimates, HomVEE reviews those estimates using WWC guidance on reviewing for Complier Average Causal Effects (CACE).[33] Under the WWC's CACE review standards, findings based on a TOT approach can receive a rating of high or moderate if they satisfy additional criteria that are not required for findings based on ITT approaches.[34]

4. Eligible outcomes

If a study uses an eligible design and analysis method, the reviewer then assesses whether the manuscript about the study includes at least one unique finding that falls into one of HomVEE's eight domains. (If any finding meeting these eligible outcomes criteria is also a subgroup finding, additional considerations apply around HomVEE's review and reporting of the subgroup finding. See Exhibit II.10 for details about HomVEE reporting on subgroups.)

a. Reporting on unique findings

Unique findings report results on a different outcome, sample or subgroup, or time period, or with a different analytic approach, than findings reported in other manuscripts about the same home visiting model.[35] HomVEE typically reviews all unique findings that authors report and for which HomVEE can calculate the statistical significance of intervention impacts. HomVEE does not consider simple transformations of analyses with the same sample, outcome, and time period to be unique findings within a study if they (1) transform findings data from frequency to a ratio (such as percentage or per thousand) or (2) transform findings data across different ratio types (such as from percentage to per thousand) because these simple transformations do not constitute a different analytic approach. In studies with such transformations, HomVEE will review the finding that is calculated as a percentage point change, because it is an intuitive measure to many readers and can be easily compared across studies for the same outcome measure.

b. Eight outcome domains

HomVEE reports only findings that can be categorized into one of HomVEE's eight outcome domains, which align with the benchmark and individual outcome domains specified in MIECHV's authorizing statute (Social Security Act, Section 511 [42 U.S.C. 711]). Because some studies follow a research sample over time, some outcomes can be assessed after early childhood. Findings that do not fall into one of these domains (Exhibit III.2) are not eligible for review and are not reported. Findings that *do* fall into

[32] Although HomVEE prefers to review ITT over TOT analyses, researchers might engage in TOT analyses to address other important research questions.
[33] See Appendix G in the WWC Procedures Handbook, Version 4.1 (U.S. Department of Education 2020a) and Section II.D. in the WWC Standards Handbook, Version 4.1 (U.S. Department of Education 2020b).
[34] These additional criteria include (1) demonstrating that any differences in outcomes in the intervention and the control groups can solely be attributed to the effects of taking up the intervention and (2) the instrumental variables included in the analyses are strong predictors of intervention take-up.
[35] Reviewers exclude non-unique findings that are re-reported in a separate manuscript. In these cases, the review simply references the other manuscript (the first or most complete one in which HomVEE encountered the finding) where HomVEE users can find those results and the review conclusions.

one of these domains are still subject to face validity and reliability requirements (see Section B.4 below). Additional detail on the outcomes in these domains is included in Appendix B.[36]

Exhibit III.2. Eligible domains and outcome examples

Domain	Examples
Child health	Measures of a child's growth, physical health, most use of health services or health care encounters (except those due to injury or ingestion), and diet or feeding.
Child development and school readiness	Child social behaviors, attachment to a parent or caregiver (as measured by observing child behavior), social-emotional or psychological development, mental and behavioral health, or cognitive and academic development. Parent or child reports of the child running away from home are also included in this domain.
Family economic self-sufficiency	Measures of a family's economic well-being, including income and earnings as well as receipt of means-tested public assistance and access to resources such as housing, food, and transportation. Family economic self-sufficiency outcomes also measure employment and educational enrollment or attainment, and other sources of support, such as child support from a noncustodial parent. Measures of the mother's partnership status (married, cohabiting, etc.) are ineligible for review. Health insurance coverage is included in this domain.
Linkages and referrals	Measures assessing whether the home visiting model has referred a family to services such as early intervention, child care, or public benefit programs.
Maternal health	Maternal health status (during or after pregnancy), including mental and behavioral health, stress, and health-related habits such as nutrition and sexual health, and measures of social support and other protective factors.
Positive parenting practices	Parent knowledge of child development, safety practices, supportive behavior and engagement with the child, promotion of learning and child development, disciplinary practices, and general parenting practices such as bedtime routines. Parent-child attachment measures that assess parental behavior are in this domain.
Reductions in child maltreatment	Measures and assessments related to child maltreatment, including evidence of substantiated child maltreatment from administrative records, and health care encounters for injuries and ingestions. Unsubstantiated reports of abuse or neglect are ineligible for review.
Reductions in juvenile delinquency, family violence, or crime	Domestic and family violence, interactions with the justice system by the mother or by a youth who received home visiting services during early childhood, or school suspensions or expulsions for one of these youth.

B. Standards for reviewing eligible designs and outcomes

After determining the eligibility of the design, analyses, and outcomes, HomVEE assesses the causal validity (rigor) of the study design reported in eligible manuscripts and applies a rating of high, moderate, or low (Exhibit III.3). HomVEE applies these ratings according to the highest rating assigned to any finding that was eligible for review in the manuscript. Most manuscripts that report study designs eligible for review by HomVEE could earn any of the three ratings, although the highest possible rating for a manuscript about a non-experimental comparison group design is moderate (Exhibit III.3). If a manuscript has no outcomes that meet all of the criteria for either the high or moderate ratings, the manuscript is rated low.

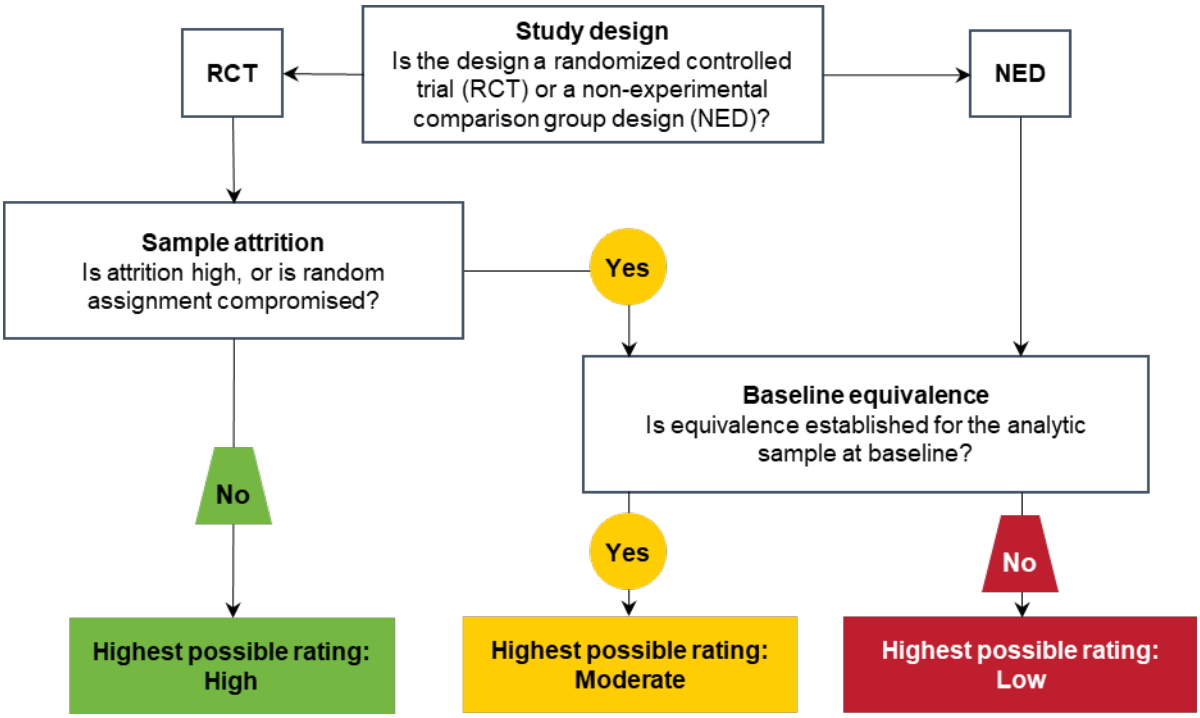
To assess the rigor of the study design, HomVEE applies standards specific to each design. HomVEE's standards for reviewing eligible randomized controlled trials and non-experimental comparison group

[36] When an assessment has scale or subscale scores that relate to several HomVEE domains, HomVEE usually places all outcomes and scales related to the assessment into a single domain. This eliminates the risk of individual subscale scores influencing the overall score, which would create unintended consequences for applying HHS criteria. An exception to this rule is assessments that have multiple scales but no overall score (such as the Protective Factors Survey). In that case, HomVEE sorts each scale into the domain to which it belongs.

designs were developed in consultation with experts and are aligned with the standards developed by the WWC. Specifically, these HomVEE standards generally are aligned with Version 4.1 of the WWC Procedures and Standards (U.S. Department of Education 2020a, 2020b), as described next in Sections B.1 (standards for RCTs), B.2 (standards on baseline equivalence), and B.3 (standards for NEDs) of this chapter.[37]

We summarize the standards for reviewing RCT and NED designs in Exhibit III.3. The remainder of this section presents details on the application of HomVEE's baseline equivalence and statistical control requirements for RCT and NED research. The standards for regression discontinuity designs and single-case designs are described in Appendix C (regression discontinuity designs) and Appendix D (single-case designs); each of those appendices consists of the WWC standards with slight wording modifications to align with terms that HomVEE uses. For example, for these two designs, HomVEE uses different terminology (high, moderate, and low) for ratings than WWC (meets standards without reservations, meets standards with reservations, does not meet standards).

Exhibit III.3. Summary of HomVEE requirements for RCTs and NEDs that do not include imputed data



Source: HomVEE standards for RCT and NED research.

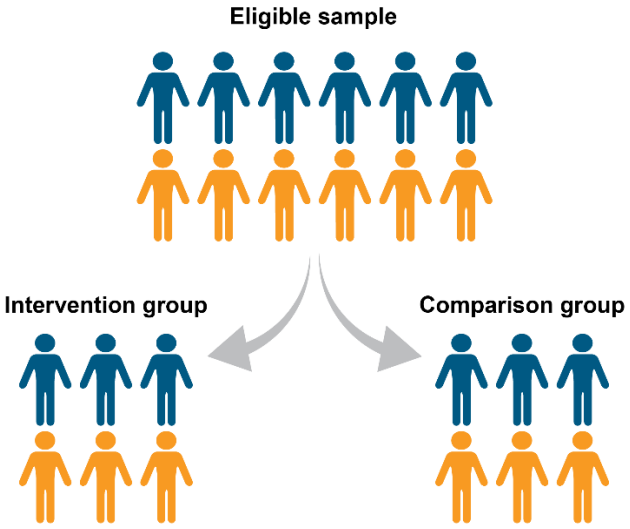
[a] The figure depicts RCT and NED designs because their standards are linked. RDD and SCD design research is not depicted here (see Appendices C and D), nor is cluster RCT and cluster NED research, because HomVEE follows a different, separate review process for those designs.

[37] These standards align with the WWC Procedures and Standards Version 4.1 (U.S. Department of Education 2020a; 2020b).

1. HomVEE standards for assessing the rigor of randomized controlled trials

Well-implemented RCTs can provide highly credible evidence about effectiveness because they create intervention and comparison groups that have equivalent measured and unmeasured characteristics, on average (see Chapter III, Section A.1. on eligible designs and Exhibit III.4). These studies provide strong evidence that differences in the outcomes of the intervention and comparison groups at follow-up can be attributed to the intervention rather than to preexisting differences between the groups (Shadish et al. 2002). RCTs are eligible to receive a high rating.

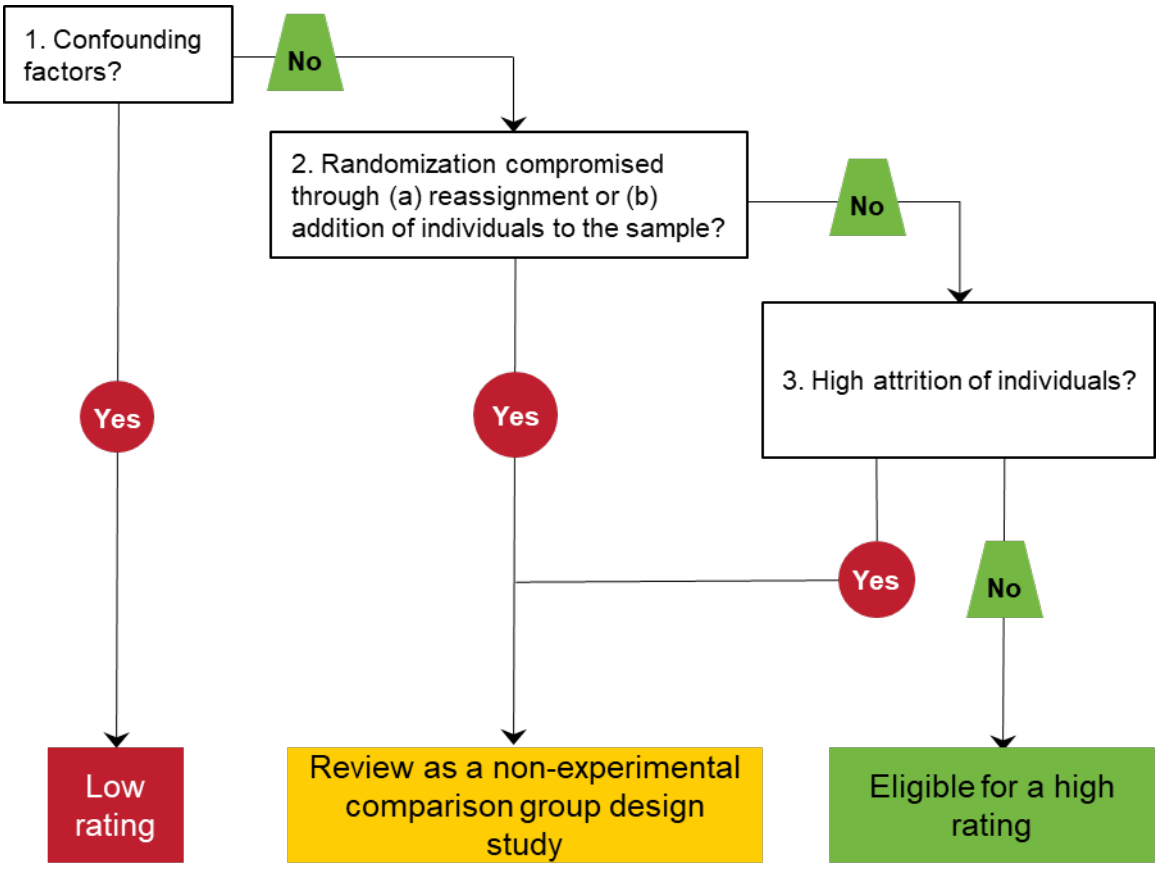
Exhibit III.4. HomVEE relies on equivalent groups to have confidence in effects of studied interventions



The HomVEE standards have several steps. First, a reviewer confirms that the author assigned sample members randomly. For a manuscript reviewed under HomVEE RCT standards, the assignment can be literally random (as in a lottery) or functionally random. For example, assigning babies with even-numbered birth dates to the intervention group and babies with odd numbers to the comparison group is a functionally random assignment process because those numbers are not expected to affect outcomes.

After confirming that the manuscript reporting the study qualifies for review as an RCT that randomly assigned individual sample members to intervention and comparison conditions, reviewers follow three additional steps to assign a manuscript rating (Exhibit III.5).

Exhibit III.5. Steps in the review process for rating randomized controlled trials with individual-level randomization



Note: To receive a rating of high, a finding must also be based on an eligible outcome that meets the validity and reliability requirements described in Section III.B.4. If findings based on imputed missing outcome data, they must meet additional requirements, as described in Appendix E.

Step 1: Are there confounding factors?

In certain cases, an element of the research design or methods lines up exactly with the intervention being tested, confounding efforts to attribute an observed effect solely to the intervention. HomVEE recognizes that models may adapt and experiment with their approach over time; thus, for HomVEE, a **confounding factor** is any observed factor that is completely aligned with either the intervention or comparison group. This means that the factor is present in only the intervention group or the comparison group, but not both.[38]

For example, if there is only one sample member in the intervention or comparison condition, there is no way to distinguish the effects of the intervention from the influence of the characteristics of that one sample member. This also would happen, for example, if one provider were assigned to all of the families in only one of the study conditions. In this case, the effect of the provider could not be separated from the effect of the intervention. (If a single agency provided home visits through multiple home visitors, HomVEE would not consider this to be a confounding factor.)

[38] Confounding factors has a more nuanced definition in the case of SCD research. See Appendix D.

A confounding factor could also arise from systematic differences in the manner (such as the measures used for each study condition) or timing of data collection from the intervention and comparison groups—for example, if program staff collected data from all participants in the intervention group, but data for the comparison group came from an administrative data set— or if researchers collected intervention group data in Year 1 and comparison group data in Year 2. 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 research design, manuscripts receive a low rating when a confounding factor is present in the study they report. Once a manuscript receives a low rating, HomVEE stops reviewing it. HomVEE reports the low rating but does not report the findings from the manuscript.

Step 2: Has random assignment been compromised?

In random assignment evaluations, deviation from the original random assignment can also bias the impact estimates. Any movement or nonrandom placement of sample members compromises the integrity of the random assignment because it could create differences in the measured or unmeasured characteristics of the intervention and comparison group members. For example, consider a study in which a program administrator reassigned families from the comparison group to the intervention group because she felt these families could greatly benefit from the intervention. Such nonrandom selection could lead to bias in the treatment effect estimates or compromise baseline equivalence (Gartin 1995).

Therefore, in order for an RCT to meet HomVEE criteria for a high rating, the analysis must randomly assign all sample members and analyze the outcomes of the intervention and comparison group members according to the group to which sample members were assigned. Sample members may not be reassigned for reasons such as contamination, noncompliance, or level of exposure. If there is evidence that random assignment has been compromised, such as by **reassignment** (the researcher moving a sample member from the intervention group to the comparison group after random assignment), the manuscript about the study is reviewed as an NED using the standards described later in this chapter. On occasion, program staff or families themselves initiate movement from the assigned intervention group into the comparison group; HomVEE does not consider this to be a reassignment problem if the research analyzes the family in the group to which the family was initially assigned.

To align its standards with those of the WWC, HomVEE checks whether the probability of assignment to each study condition was equal. If the probabilities differ (including if the probability of assignment to a group varies across blocks in a stratified random assignment framework), then the reported analysis must use one of three methods of adjustment: [39]

1. Estimate a regression model in which the covariate set includes dummy variables that differentiate subsamples with different assignment probabilities
2. Estimate impacts separately for subsamples with different assignment probabilities and average the subsample-specific impacts
3. Use inverse probability weights

[39] This criterion for assignment probability is from WWC Standards Version 4.1 (see pp. 5-6 of the WWC Standards Handbook, Version 4.1 (U.S. Department of Education 2020b).

If the manuscript text suggests authors used varying probabilities of random assignment, but does not report on or adjust for differing probabilities of being assigned to the intervention group, it is not eligible for a high rating and is reviewed using the process for NEDs.

HomVEE does not require authors to discuss why randomization was compromised, but authors usually explain why sample members were excluded from the analytic sample or why there was reassignment from one group to the other. The presence of noncompliers (individuals participating in the activities or services of the group to which they were not originally assigned) are not considered to compromise the randomization if the analyses included individuals in the groups to which they were originally assigned. In addition, if the information provided in a manuscript does not enable reviewers to clearly assess whether randomization was compromised, HomVEE will issue an author query to ask for information that would enable reviewers to appropriately assess the possible threat to the randomization.

Step 3: Is there high attrition?

Attrition happens when outcome data are missing for some members of the intervention or comparison groups. Attrition can occur because sample members do not respond to surveys or are missing from administrative data sets, or it can occur for some other reason.

In random assignment evaluations, attrition can bias the impact estimates by creating differences in the characteristics of the intervention and comparison groups, even if these groups were equivalent at the time of random assignment. If sample members in the intervention and comparison groups who remain in the study at follow-up had different characteristics at the time of random assignment, outcomes could differ even in the absence of treatment (Shadish et al. 2002) (Exhibit III.6). **So, "moderate" is the highest rating that a manuscript reporting results of an RCT with high attrition can receive.** [40]

For example, if less-motivated sample members in the intervention group failed to respond to the follow-up survey, the outcome data for the intervention group would be based on fewer unmotivated individuals than the outcome data for the comparison group. Similarly, if highly motivated sample members in the comparison group were frustrated by a lack of services and did not respond to the follow-up survey, the outcome data for the comparison group would be based on fewer motivated individuals than the outcome data for the intervention group. After this attrition, more-motivated individuals would be disproportionately represented in the intervention group, and less-motivated individuals would be disproportionately represented in the comparison group. If motivation were associated with the study outcomes, the findings would be biased and would not accurately reflect the effect of the home visiting services.

To determine attrition, HomVEE calculates the percentage of the randomized sample who did not have outcome data and were not included in the study analyses, both overall and differentially. (Additional requirements apply for cluster RCTs, which are described in Section C.1 of this chapter.)

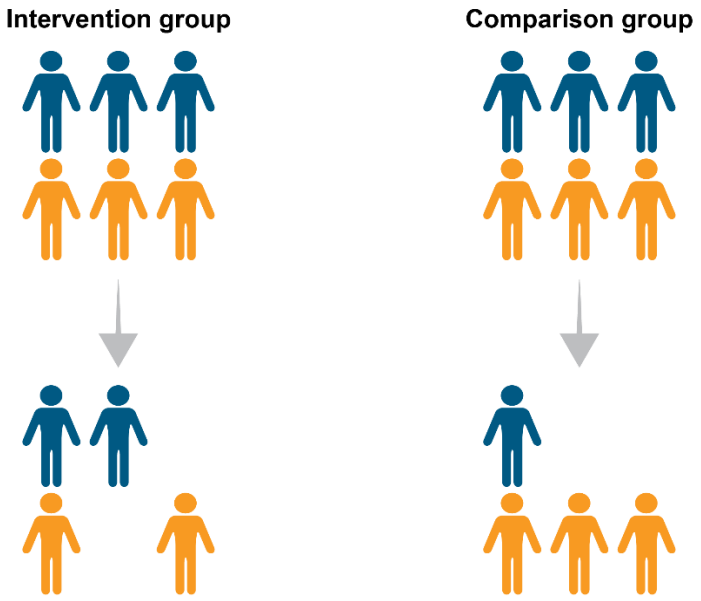
- **Overall attrition** is the combined loss of data for any sample member from either the intervention or comparison group. For example, if 100 individuals were randomly assigned (50 to the intervention group and 50 to the comparison group), and 25 did not respond to a follow-up survey, overall attrition

[40] In some designs, there is baseline missing data. HomVEE has discretion in determining whether the extent of missing baseline data poses a risk of bias. If authors use an imputation approach to handle the missing baseline data, the design will be reviewed following the standards described in Appendix E.

for the follow-up survey would be (100 – 75)/100 = 25 percent. This is the calculation regardless of whether those individuals are part of the intervention or comparison group.

- **Differential attrition** refers to the difference in the rate of attrition between the intervention and comparison groups. Consider the example above, where overall attrition is 25 percent. If 15 of the 25 individuals who did not respond to the follow-up survey were in the intervention group, then intervention group attrition is 30 percent (15/50 = .30). That leaves 10 in the comparison group who did not respond, meaning that comparison group attrition is 20 percent (10/50 = .20). The differential attrition—30 percent intervention group attrition minus 20 percent comparison group attrition—is 10 percentage points.

Exhibit III.6. After attrition, there may be differences between the intervention and comparison groups



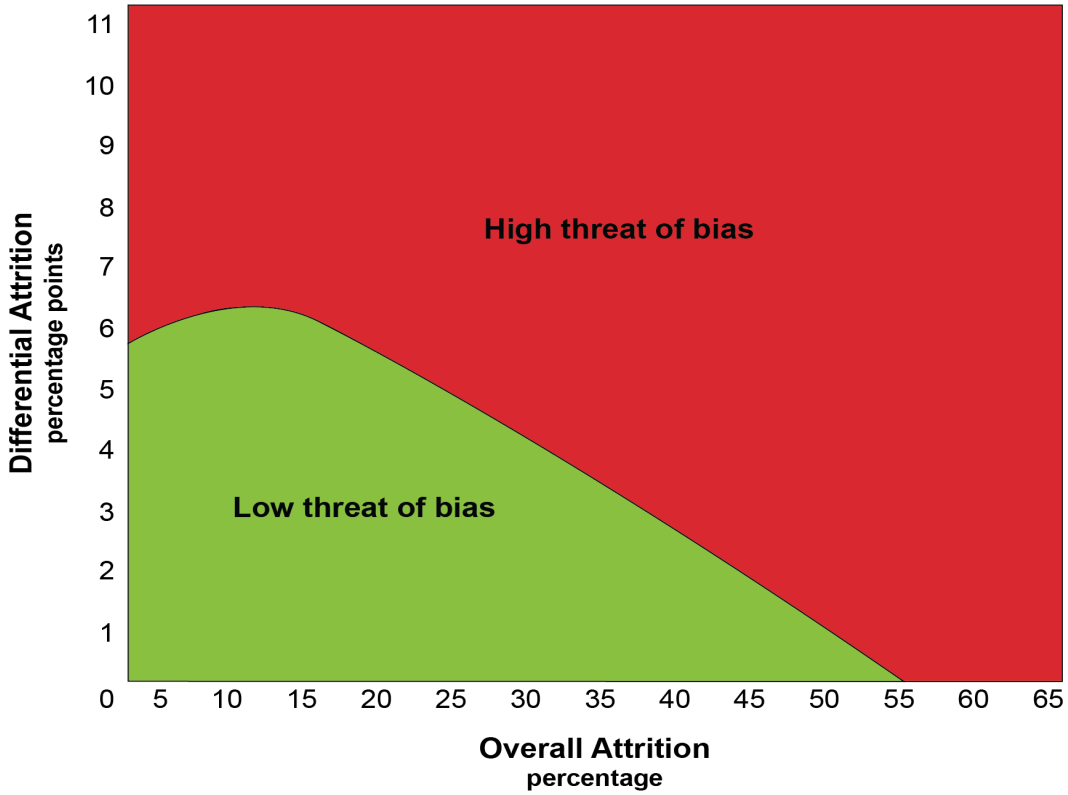
The HomVEE review uses the WWC boundary for attrition. The WWC boundary for attrition is transparent and empirically based, taking into account both overall attrition (the percentage of study participants lost from the total study sample) and differential attrition (the difference in attrition rates between the intervention group and the comparison group). It recognizes an important trade-off between overall and differential attrition—namely, that 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. (See Section II.A.2. Assessing sample attrition, in the WWC Standards Handbook, Version 4.1 [U.S. Department of Education 2020b].) **Sample members with missing and then imputed data are considered to be missing when computing attrition.** (See Chapter III, Section D, as well as Appendix E for more information on how HomVEE reviews manuscripts about studies with missing data.)

In alignment with Version 4.1 of the WWC Standards (U.S. Department of Education 2020b), some types of sample loss will not count as attrition in HomVEE. First, losing sample members after random assignment because of acts of nature such as hurricanes, fires, or the COVID-19 pandemic is not considered attrition if the loss affects the intervention and comparison conditions in the same way.

However, because there is no reason to believe that a crisis would always create a greater sample loss in either the intervention or the comparison groups, if the sample loss due to an act of nature was concentrated in one of the conditions–that is, if the difference in sample loss between the home visiting intervention group and the comparison group, based on the differential attrition rate, is substantial enough to constitute high attrition (see Exhibits III.7 and IIII.8)–then the sample loss is considered attrition. Second, when researchers exclude a subsample of the randomly assigned sample from their analysis, HomVEE does not consider that excluded subsample to constitute attrition if (1) the subsample was randomly selected or (2) the subsampling was based on characteristics that were clearly determined before the start of the intervention and applied consistently across the intervention and comparison conditions.

The WWC attrition boundary classifies research as having either "high" or "low" attrition based on a combination of overall and differential attrition (Exhibit III.7).[41] For each overall attrition rate, Exhibit III.8 shows the highest differential attrition rate allowable to still be considered "low attrition" by HomVEE.

Exhibit III.7. Overall and differential attrition levels that result in high or low attrition



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.

[41] The WWC created two possible attrition boundaries in Version 4.1, optimistic and cautious, and review team leaders select the one that makes the most sense for their review based on whether the intervention being studied is likely to affect attrition. HomVEE has adopted the cautious boundary, because home visiting may affect families' choices about continuing to participate in the study. (Earlier versions of WWC standards refer to the boundaries as liberal and conservative.)

Exhibit III.8. Highest differential attrition rate for a sample to maintain low attrition, by overall attrition rate

Overall	Differential	Overall	Differential	Overall	Differential
0	5.7	22	5.2	44	2.0
1	5.8	23	5.1	45	1.8
2	5.9	24	4.9	46	1.6
3	5.9	25	4.8	47	1.5
4	6.0	26	4.7	48	1.3
5	6.1	27	4.5	49	1.2
6	6.2	28	4.4	50	1.0
7	6.3	29	4.3	51	0.9
8	6.3	30	4.1	52	0.7
9	6.3	31	4.0	53	0.6
10	6.3	32	3.8	54	0.4
11	6.2	33	3.6	55	0.3
12	6.2	34	3.5	56	0.2
13	6.1	35	3.3	57	0
14	6.0	36	3.2	58	-
15	5.9	37	3.1	59	-
16	5.9	38	2.9	60	-
17	5.8	39	2.8	61	-
18	5.7	40	2.6	62	-
19	5.5	41	2.5	63	-
20	5.4	42	2.3	64	-
21	5.3	43	2.1	65	-

Source: WWC Standards Handbook, Version 4.1 (U.S. Department of Education 2020b), Section II.A.2, Assessing sample attrition.

Manuscripts about random assignment studies that have at least one outcome that meets the standard for low attrition are considered for the high rating. RCTs with high attrition on all eligible outcomes and in which the findings are not based on imputed outcome data are reviewed as an NED (discussed after the next section, in Section B.3) due to the bias that attrition may cause, and therefore the highest rating they may receive is moderate (if they meet the other criteria for that rating). RCTs in which the analyses use imputed outcome data (and therefore can include the full randomized sample) are reviewed following a different review process, described in Appendix E.

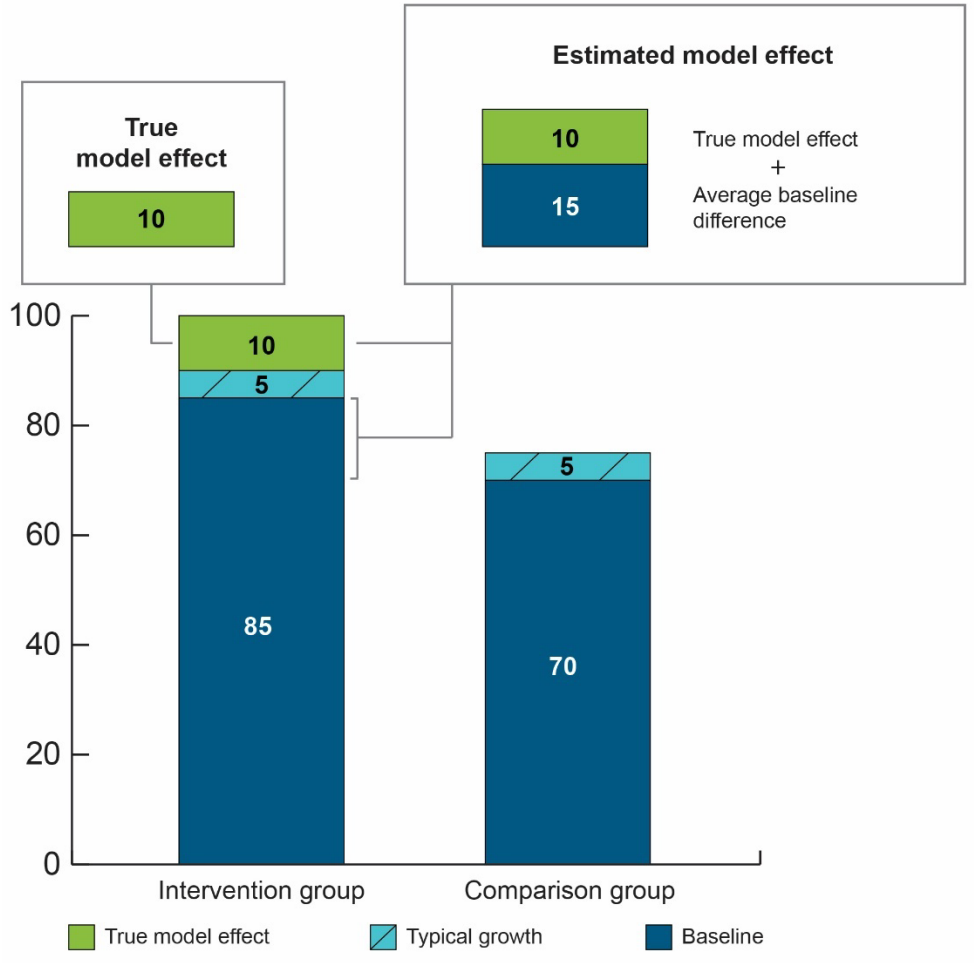
2. HomVEE standards for baseline equivalence

HomVEE standards for baseline equivalence apply to RCTs with high attrition and in which the analyses do not use imputed outcome data, RCTs with compromised random assignment, and to NEDs.

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

Bias can occur when the intervention and comparison groups in the analytic sample (that is, the sample members actually included in the analysis, which may differ from those assigned to each group) differ at baseline. For example, Exhibit III.9 shows that baseline productive vocabularies of toddlers in the intervention group (85 points) are larger, on average, than the corresponding vocabularies of those in the comparison group (70 points). Six months later, after services from the home visiting model have ended, both groups have improved simply because the children got older (5 points). After this period, the intervention group has also improved from participating in the home visiting model (10 extra points). If researchers measure outcomes at six months, there is an estimated effect of a 25-point difference between toddlers in the intervention group (85 + 5 + 10 = 100 points) and toddlers in the comparison group (70 + 5 = 75 points). The initial differences in the two groups led to a biased estimate of model effectiveness, because part of the estimated difference is due to the initial (baseline) underlying differences (15 points), and the other part (10 points) constitutes the "true" effect of the home visiting model (an unobserved effect).

Exhibit III.9. Baseline differences lead to biased estimates



a. What is baseline equivalence?

Equivalence means that the intervention and comparison groups are similar on specified characteristics. The equivalence between the intervention and comparison groups must be established at **baseline**—that

is, before the intervention being studied is provided to the intervention group.[42] In the next section, we describe HomVEE's baseline equivalence requirements, which in some instances include controlling for baseline characteristics in the impact analyses.

Establishing baseline equivalence supports conclusions that the treatment, and not preexisting differences, led to any observed difference in outcomes (Shadish et al. 2002). Characteristics may change over time, sometimes as a result of the home visiting model itself. For example, a home visiting model may offer services to support a family's economic well-being, which could affect measures of socioeconomic status (SES). So, for example, if a home visiting model starts in pregnancy, then SES data from the time of the birth (such as whether Medicaid paid for the hospital stay, or from vital statistics collected to develop a birth certificate) are not actually a baseline measure of SES.

RCTs create intervention and comparison groups that are expected to be equivalent across both measured and unmeasured baseline characteristics because the assignment to the groups was random. However, when there is high attrition, or the design has been compromised (with reassignment or addition of sample members) in an RCT, the intervention and comparison groups in the sample of families that contributes outcome data for the analysis (that is, the analytic sample) might no longer be similar because they are not the same groups that were randomly assigned to each condition at baseline.

In the case of NEDs, the assignment to the intervention and comparison groups is not random, so there could be differences in the key baseline characteristics of the intervention and comparison groups. **For those reasons, NEDs and RCTs with high attrition) or a compromised design cannot be considered for a high rating, but they may be considered for a moderate rating if at least one of their findings meets HomVEE's requirements for baseline equivalence on (1) race/ethnicity, (2) socioeconomic status (SES) and (3) baseline measures of outcomes (when feasible).** These baseline equivalence requirements are described next, and the process to assign ratings to NEDs and RCTs with high attrition or a compromised design (and that do not use imputed outcome data) is described in the section after that, in Section B.3.

b. Assessing and satisfying baseline equivalence requirements for NEDs and RCTs with high attrition (but no imputed outcome data) or compromised designs

HomVEE reviewers assess whether the study groups were equivalent, at baseline, on the analytic sample. Equivalence must be established on the analytic sample (the families that contribute to the outcome data used in the analysis); equivalence on the sample assigned to the intervention and comparison conditions at the beginning of the study will not satisfy this criterion if the analytic sample is smaller than the sample assigned to intervention and comparison conditions. If sample members drop out of the study or are not assessed at follow-up, groups that began as equivalent might have quite different compositions by the follow-up. For HomVEE, manuscripts must demonstrate baseline equivalence for the sample included in the follow-up impact analysis.[43]

[42] If the data collection timing is not clear, then HomVEE queries the authors for clarification. Race/ethnicity is a time-invariant characteristic, so HomVEE accepts any timing of race/ethnicity measures when establishing baseline equivalence.

[43] Typically, HomVEE prefers that authors establish that all analytic samples are equivalent. However, HomVEE also aims to reduce unnecessary burden on authors to establish equivalence for the sample on every finding in cases of item-level attrition where equivalence for a near-identical sample on another finding has already been established. HomVEE uses a guideline of 10 percent overall attrition and 2 percent differential attrition, relative to an equivalent

In alignment with Version 4.1 of the WWC Standards, **to verify baseline equivalence in a specified characteristic, HomVEE reviewers will look at the effect size (ES)[44] of the difference between the intervention and comparison groups in the specified baseline characteristic.** That is, HomVEE uses effect size, computed as the absolute value of the difference between intervention and comparison groups in standard deviation units, to verify baseline equivalence. HomVEE requires the following to be true for research to demonstrate baseline equivalence for a specified characteristic:[45]

- A baseline effect size less than or equal to 0.05 meets the baseline equivalence requirement and requires no statistical adjustment.
- For a baseline effect size that is greater than 0.05 and less than or equal to 0.25, an acceptable statistical adjustment for the baseline characteristic is required to meet the baseline equivalence requirement.
- For a baseline effect size greater than 0.25, HomVEE considers the intervention and comparison groups to be nonequivalent; that is, the intervention and comparison groups do not meet the baseline equivalence requirement for the specified characteristic.

A **statistical control or statistical adjustment** is a method researchers use to include the baseline measures in a statistical model. When statistical adjustment is necessary to establish baseline equivalence, HomVEE requires that the baseline measures required for establishing equivalence be included in the analysis at the same level of the unit of analysis so that the analysis accounts for the correlation between the baseline measure and the outcome.[46] Several techniques satisfy this requirement, such as regression adjustment and analysis of covariance (see Exhibit III.10).

For establishing baseline equivalence on outcomes, HomVEE considers additional methods of adjustment acceptable whenever the outcome measures are the same at baseline as at follow-up. These methods include gain scores and difference-in-difference adjustments. These methods are acceptable when the following two conditions are satisfied (WWC Standards Handbook, Version 4.1):

1. Authors used the same units to measure the outcome measures at baseline and follow-up. This condition is not satisfied when (a) the authors administered different assessments at baseline and follow-up, or, (b) the measures at baseline and follow-up are the same, but different subscales or scoring procedures were used to construct the measure.

---

sample, to prompt a query for equivalence on a related sample. This guideline falls well within both sides of the attrition boundary (see Exhibit III.7), giving HomVEE strong confidence that the samples will not be markedly different. For example, if authors established equivalence for a parent interview sample consisting of 100 families each in the intervention and comparison conditions at 12 months, but 4 parents from the treatment and 7 from the comparison condition did not answer one of the questions, HomVEE may decide to proceed with reviewing that finding with minimal missing item-level data without sending additional questions to the author.

[44] To limit bias that can arise from differences in the treatment and comparison group units used to measure the effect of a home visiting model on outcomes, the groups must appear similar on the relevant baseline characteristics that are thought to be related to the outcomes. This balance must be shown to be achieved using the observed data in the sample, and these differences can be measured using the effect size (Ho, Imai, King, & Stuart 2007; Imai, King, & Stuart 2008). HomVEE will use the same formulas the WWC uses to calculate effect sizes (differences in standard deviations), for both continuous and dichotomous variables, as described in Chapter VI of the WWC Procedures Handbook, Version 4.1 (U.S. Department of Education 2020a).

[45] These requirements are outlined in the WWC Standards Handbook, Version 4.1 (U.S. Department of Education 2020b).

[46] This requirement is in alignment with Version 4.1 of the WWC procedures and standards (U.S. Department of Education 2020a; 2020b).

2. The baseline measure of the outcome has a strong relationship with the follow-up measure of the outcome. That is, the correlation of the baseline and follow-up measures of the outcome is 0.60 or higher. Authors must have estimated this correlation using the data from the analysis under review.

When a statistical adjustment is not required (in the case of low-attrition RCTs or baseline differences less than or equal to 0.05 standard deviations), authors can use approaches other than those HomVEE accepts to adjust their analyses for the purpose of satisfying the statistical adjustment requirement. Additionally, although HomVEE requires statistical adjustments in limited circumstances and only for certain specified characteristics, authors may adjust for all available baseline data in their analyses.

Exhibit III.10. Statistical adjustment methods accepted by HomVEE

Acceptable analytic methods to adjust for baseline differences:

- Regression adjustments
- Analysis of covariance (ANCOVA) or multivariate analysis of covariance (MANCOVA)
- Estimating impacts only for groups defined at baseline (for example: ever had a baby, never had a baby)
- Repeated measures analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA) (these approaches to modeling repeated measures research are also subject to other requirements; see Chapter III, Section C.2 on repeated measures analyses)
- Growth curve modeling (this approach to modeling repeated measures research is also subject to other requirements; see Chapter III, Section C.2 on repeated measures analyses)

Acceptable methods if baseline and follow-up measure of outcome are the same and have a strong relationship to each other

- Gain or change scores (pre-post differences)
- Difference-in-difference adjustments
- Fixed effects for individuals

Note: These methods align with WWC Version 4.1 Handbook. In addition, HomVEE specifies that the following three methods of adjustment are acceptable: (1) repeated measures ANOVA or MANOVA; (2) estimating impacts only for groups defined at baseline (for example, ever had a baby versus never had a baby); and (3) growth curve modeling. The WWC Version 4.1 Handbook does not mention these three methods.

HomVEE will also consider the following when assessing baseline equivalence: [47]

- Baseline data that include imputed data can be used to demonstrate baseline equivalence of the analytic sample, but the process is different than it is when the baseline data do not include imputed data. If the baseline data include imputed data, HomVEE will first estimate how large the baseline difference (in standard deviation units) between intervention and comparison groups might be under different assumptions about how the missing data are related to measured and unmeasured factors. Then, HomVEE uses the largest of those estimates in absolute value as the effect size for assessing baseline equivalence. This process is aligned with Version 4.1 of the WWC Standards.

[47] All of these considerations align with Version 4.1 of the WWC standards.

- In RCT designs in which random assignment occurred at the individual/family unit level (so the unit of assignment and the unit of analysis are the same: the individuals/family units), the measures used to establish baseline equivalence of individuals/family units must be at the individual/family unit level. That is, HomVEE will not accept measures at an aggregate level (such as community- or county-level measures, for example) to show baseline equivalence of the sample of individuals/family units in an RCT in which individual or family units are assigned to intervention or comparison groups.
- If the impact analyses use weights, then the baseline means must be calculated using the same weights.
- If the study conducted random assignment within blocks or strata, and the analyses included dummy variables that differentiate these blocks or strata, then these same dummy variables can be used to adjust the baseline means.

If the study provides no information that indicates demographic measures or baseline measures of outcomes were collected, HomVEE assumes that the groups were *not* equivalent at baseline and does not query authors for further information.

i. Establishing equivalence on demographics: Race, ethnicity, and socioeconomic status

HomVEE's definition of baseline equivalence applies to race/ethnicity, socioeconomic status, and baseline measures of outcomes. HomVEE requires baseline equivalence on the demographic characteristics because they may be related to the outcome domains that are the focus of the 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 or ethnicity of the participant. For example, research shows that the birth outcomes of various race/ethnicity groups are significantly different from each other (MacDorman 2011).

SES can be measured in multiple ways. HomVEE prefers 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). However, HomVEE also accepts means-tested assistance measures (see Exhibit III.11) because they are closely tied to the HomVEE preferred measures of SES (income, earnings, and poverty level). These measures are common indicators of SES and are relevant to the population served by home visiting models.

Exhibit III.11. HomVEE baseline equivalence requirements

Baseline category	Accepted measures
Race/ethnicity	HomVEE generally accepts author-reported race/ethnicity categories. Information on nationality and citizenship does not satisfy HomVEE's requirements for race/ethnicity. A manuscript that indicates that the intervention and comparison groups are racially or ethnically homogenous (for example, a sample that exclusively consists of Hispanic mothers) satisfies HomVEE's baseline equivalence requirement for race/ethnicity. Baseline equivalence must be established across all race/ethnicity categories reported by the authors. Although HomVEE would prefer authors to report race/ethnicity for parents and for children in the sample, HomVEE will accept the race/ethnicity measure of one generation of the family as a proxy for the other generation if the parent and child are biologically related and if the author didn't measure race/ethnicity for both generations.[a]
Socioeconomic status	Preferred: One of the following economic well-being measures shows equivalence: • Income • Earnings • Poverty levels according to federal thresholds • Maternal education[b] **OR** Alternative: At least two of the following secondary measures show equivalence: • Means-tested assistance measures, such as Temporary Assistance for Needy Families (TANF), Supplemental Nutrition Assistance Program (SNAP) receipt, Medicaid receipt, or the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) • Employment of at least one household member Other measures of SES are rarely accepted, especially if they use subjective thresholds that vary with cultural norms—one example is measures of overcrowding in the home (Myers et al. 1996). Rarely, HomVEE may accept other secondary SES measures, such as in research conducted outside of the United States or when constraints of the research context make typical secondary measures difficult to assess (for example, an indicator of electricity or indoor plumbing access in research conducted in aboriginal communities). If a manuscript reports one of the specific economic well-being measures and two or more of the secondary measures above, HomVEE will rely upon the well-being measure to determine baseline equivalence.
Outcomes measured at baseline	If such measures are age or developmentally appropriate to collect at baseline (see Appendix B), researchers should provide data from a baseline measure that is the same as or similar to the outcome measure (see Section B.2.b.ii on establishing equivalence on baseline measures of outcomes).

[a] HomVEE recognizes that there is a limitation in allowing the race/ethnicity of one generation of a family to proxy for the other generation, especially for children in mixed race households. However, the chance of that introducing bias (differential incidence across study arms) is assumed to be low.

[b] Original HomVEE standards treated maternal education as a secondary measure of SES. In the Version 2 Handbook, maternal education was updated to be treated as a primary measure of SES based on subject matter expert recommendations and research demonstrating that maternal education is a robust proxy for socioeconomic status and a key predictor of family and young children's outcomes (for example, Jackson et al., 2017; Patra et al., 2016; Schochet et al., 2020).

ii. Establishing equivalence on baseline measures of outcomes

For home visiting interventions that serve pregnant women and families with young children, it is sometimes impossible or inappropriate to examine the same variables at baseline and follow-up. We present our criteria for two scenarios in this section.

Scenario 1: *Measures assessing variables identical or sufficiently similar to the outcomes of interest are* ***assessable at baseline*** *(that is, if they are age- or developmentally appropriate to collect).* When possible, baseline equivalence should be established on the outcomes of interest. HomVEE will consult with subject matter experts and HHS to determine whether baseline variables are sufficiently similar to outcome variables.

Scenario 2: *Measures of outcomes of interest are not assessable at baseline.* Sometimes, measures of the outcomes of interest cannot be assessed at baseline. For example, researchers cannot collect baseline cognitive skills for a child when program services start prenatally. Therefore, when assessing equivalence, HomVEE reviewers also consider whether the measure was assessable at baseline when deciding whether the study must demonstrate equivalence on the measure.[48] On the measures that are not assessable at baseline, HomVEE assumes equivalence (but still checks race/ethnicity and SES equivalence). Appendix B lists, for each HomVEE outcome domain, measures or groups of measures that are and are not considered assessable at baseline. This list is based on measures already encountered by HomVEE reviewers, and it represents guiding principles for how additional measures may be handled if they are reviewed in the future.

In addition to these requirements, HomVEE, in consultation with subject matter experts, has the discretion to determine other cases where baseline equivalence is insufficiently demonstrated. For example, some measures of race/ethnicity that combine a wide range of responses (such as a measure that groups all non-White persons into one category and all White persons into another) may not be appropriate to demonstrate baseline equivalence. Also, not establishing baseline equivalence on variables other than race/ethnicity, socioeconomic status, and baseline outcomes (that is, on variables not required to establish baseline equivalence for the purposes of the HomVEE review) could be an indicator that the intervention and comparison groups in the analytic sample are not equivalent. In these cases, project leaders have the discretion to determine whether baseline equivalence has been sufficiently demonstrated.

3. Non-experimental comparison group designs (NEDs) and RCTs with high attrition (but no imputed outcome data) or a compromised design

The highest rating that NEDs with an external (that is, non-overlapping) comparison group can achieve is moderate. In such studies, participants are sorted into the study arms through a process other than random assignment and must be in no more than one group; therefore, even if the treatment and comparison groups are well matched based on observed characteristics, they may still differ on unmeasured characteristics. It is thus impossible to rule out the possibility that the findings are attributable to unmeasured group differences. Similarly, in RCTs with high attrition that do not use imputed outcome data (and therefore cannot include the full randomized sample), or in RCTs with a compromised design, it is also impossible to rule out that the findings are due to the compositional changes in the intervention and comparison groups and not to the intervention. Therefore, HomVEE reviews NEDs and RCTs with high attrition (and no imputed outcome data) or a compromised design with the same standards.

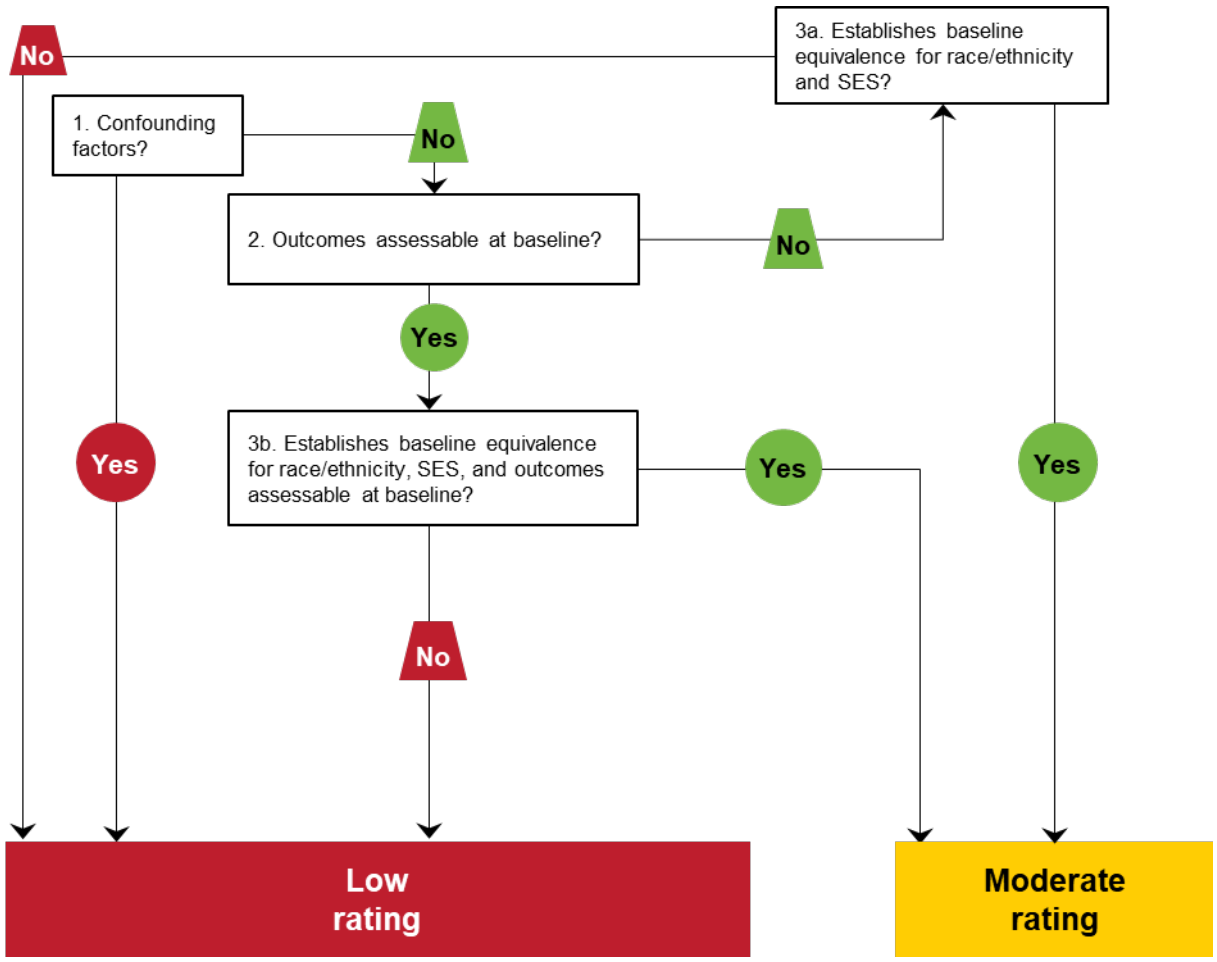
Designs without a comparison group (for example, pre-post designs) offer no way to assess what the sample's outcomes would have been in the absence of the intervention. These study designs cannot rule out that the changes were caused by, for example, history (an event besides the intervention that could have produced the observed outcome) or maturation (participants' natural changes over time could have

[48] HomVEE considers an outcome to be assessable at baseline only if that outcome was assessable for the **entire sample** at baseline (or, if the analysis examines a subgroup, if the outcome was assessable for the entire subgroup at baseline).

produced the outcome) (Shadish et al. 2002). Therefore, manuscripts about studies with these designs rate low.

NEDs and RCTs with high attrition (but in which the analyses do not use imputed data) or a compromised design are eligible for a rating of moderate if there are no confounding factors in the design and if they meet the baseline equivalence requirement described in Chapter III, Section B.2. Specifically, HomVEE uses a four-step process for rating findings from NEDs and RCTs with high attrition (but without imputed outcome data) or a compromised design (Exhibit III.12).

Exhibit III.12. Steps in the review process for rating findings from NEDs, RCTs with high attrition (but no imputed outcome data), or RCTs with a compromised design



SES = Socioeconomic status.

Note: Analyses need to adjust for race/ethnicity and socioeconomic status if the difference in standard deviations between the intervention and comparison groups in these characteristics is equal or greater than 0.05 and less than 0.25.

Step 1: Are there confounding factors?

HomVEE uses the same approach to confounding factors within NEDs and RCTs with high attrition or a compromised design (but that do not include imputed outcome data) as it does for RCTs with individual- or family-level randomization (Chapter III, Section B.1 on standards for reviewing RCTs). If a

confounding factor is identified within a study, it automatically receives a low rating and is not reviewed further.

Step 2: Are outcomes assessable at baseline?

If there are no confounding factors, HomVEE reviewers check whether outcomes are assessable at baseline. They are guided by the list of measures or groups of measures that are and are not considered assessable at baseline, by outcome domain. The list is in Appendix B. If outcome measures are not assessable at baseline, HomVEE considers the intervention and comparison groups in the analytic sample to be equivalent at baseline in such measures, and the review continues to Step 3. If outcome measures are assessable at baseline, HomVEE reviewers proceed to Step 4.

Step 3: Is baseline equivalence established for race/ethnicity and socioeconomic characteristics?

If no outcomes are assessable at baseline, reviewers check whether baseline equivalence is established for race/ethnicity and socioeconomic characteristics. To receive a moderate rating, NEDs, RCTs with high attrition (but that do not use imputed outcome data), or RCTs with compromised designs must meet HomVEE's baseline equivalence requirements on race/ethnicity and socioeconomic characteristics, as specified in Chapter III, Section B.2 (HomVEE's baseline equivalence requirements). If this baseline requirement is met, the findings receive a moderate rating. If the baseline requirement is not satisfied, then the findings rate low.

Step 4: Is baseline equivalence established for race/ethnicity, socioeconomic characteristics, and outcomes assessable at baseline?

For any outcomes that are assessable at baseline, NEDs, RCTs with high attrition (but that do not use imputed outcome data), or RCTs with compromised designs must satisfy HomVEE's baseline equivalence requirements, on race/ethnicity, socioeconomic characteristics, *and baseline measures of outcomes that are feasible to measure at baseline*. These baseline equivalence requirements are described in detail in Chapter III, Section B.2. If baseline equivalence is established, then the findings on outcomes that were assessable at baseline receive a moderate rating. If the baseline requirement is not satisfied, then those findings receive a low rating.

Steps 3 and 4 apply to NEDs, RCTs with high attrition (but that do not use imputed outcome data), or RCTs with compromised designs that do not have missing or imputed baseline data on race/ethnicity, socioeconomic characteristics, and outcomes assessable at baseline. The process to review any RCTs or NEDs that have missing or imputed baseline data on any baseline measures required for establishing baseline equivalence is described in Appendix E.

4. Face validity and reliability requirements for measures

For findings to be eligible for a high or a moderate rating, they must measure the effect of an intervention on outcome measures that demonstrate face validity (the outcome measures actually measure the concepts they seek to measure) and reliability (the outcomes measure concepts accurately). These measurement standards are aligned with Version 4.1 of the WWC Standards (U.S. Department of Education 2020b).

a. Face validity

To demonstrate face validity, an outcome measure must have a detailed enough description within the study to enable HomVEE's reviewers to determine that the measure is (1) clearly defined and (2) the content the measure assesses aligns with its definition. For example, a measure described as an indicator of a child's health that actually measures a child's behavior does not have face validity. To verify the face validity of an outcome measure, HomVEE reviewers check the elements listed in Exhibit III.13. Findings based on outcome measures that do not meet the face validity standard will rate low.

Exhibit III.13. Elements HomVEE examines when assessing whether an outcome measure demonstrates face validity

Item to check	Description of the requirement
Is the outcome measure clearly described?	Authors must provide a description of the measure that allows reviewers to understand the source of the measure or how it was created, and what it intends to measure. A clear description of the measure allows reviewers to verify that the measure (1) belongs in at *least one* of the eight domains of interest for HomVEE, and (2) has a clear connection with the construct it claims to measure (that is, the measure is applied as it was designed to be and is constructed with items from only the domain or domains it is intended to measure).
What type of measure is this outcome measure?	
Is the measure a standardized measure?	If the measure is a standardized measure, HomVEE prefers that authors provide citations to support the description of the measure. If no citations are provided for standardized measures, HomVEE will request them through an author query if needed.
Is the measure a modified version of a standardized measure?	For a modified standardized measure, authors must clearly describe which items were dropped from a standardized measure and provide a clear interpretation for the modified standardized measure. HomVEE prefers that authors provide citations to support the description of the measure. If no citations are provided, HomVEE will request them through an author query.
Is the measure a new measure created for the study described in the manuscript under review?	For new measures, authors must provide a clear description of how they constructed the measure. Authors must also provide citations to support the description of the measure, if applicable.

If a manuscript does not provide citations (and the review team determines they are needed) and/or there is not enough information in the description of the measure provided in the manuscript under review, HomVEE will issue an author query to ask for information that will allow reviewers to assess whether the measure is clearly defined, including a detailed description of (1) the items included in the measure, (2) the methodology used to construct the measure, and (3) interpretation of the measure. Reviewers will also ask for citations supporting the methods used to construct the measure and supporting its interpretation, if needed. In addition, whenever it is not clear whether a measure meets the validity requirement (even after receiving a response to the author query), HomVEE reviewers will consult with project leaders, and the project leaders will consult with subject matter experts and with ACF about the validity of measures.

b. Reliability

HomVEE reviewers will apply reliability standards to all outcome measures that are within one of HomVEE's eight outcome domains. Findings based on outcome measures and/or baseline measures that do not meet the reliability standards will rate low.

Some measures are not appropriate to validate with psychometric tests. HomVEE will assume that the following measures are reliable: (1) administrative records obtained from child welfare or other social service agencies, hospitals or clinics, and schools; (2) demographic characteristics; and (3) medical or physical tests.

Otherwise, to demonstrate reliability, outcome measures must meet at least one of the following standards:[49]

- Internal consistency (such as Cronbach's alpha) of 0.50 or higher.
- Test-retest reliability of 0.40 or higher.
- Inter-rater reliability (as indicated by percentage agreement, correlation, or kappa) of 0.50 or higher.

HomVEE reviewers will prioritize reliability statistics on the sample of participants in the manuscript under review, but will also consider statistics from test manuals or studies of the psychometric properties of the measures. The review team may ask authors to provide additional information about the reliability of their measures.

C. Other analysis methods

This section describes HomVEE's standards for reviewing the following designs and analytical approaches: (1) cluster RCTs and NEDs, (2) repeated measures analyses, and (3) structural equation models.

1. Cluster RCTs and NEDs

HomVEE considers designs to be cluster RCTs or NEDs whenever family units are assigned (randomly in the case of RCTs) to the intervention or comparison conditions as groups (or clusters), such as a neighborhood, ZIP code, or county.

In research that involves random assignment of clusters but uses data from family units to estimate impacts, HomVEE assesses whether the findings from cluster RCTs and NEDs can be credibly attributed to the intervention only, or whether changes in the composition of family units in the sample could have also affected the findings. For example, the composition of family units in the sample changes if (1) researchers move family units from the condition (intervention or comparison) to which they were originally assigned into the other condition, and/or (2) there is considerable nonresponse from family units at the time of the follow-up assessment when outcomes are measured. Findings receive a rating of high only when it is possible to rule out that compositional changes in the sample influenced the findings.

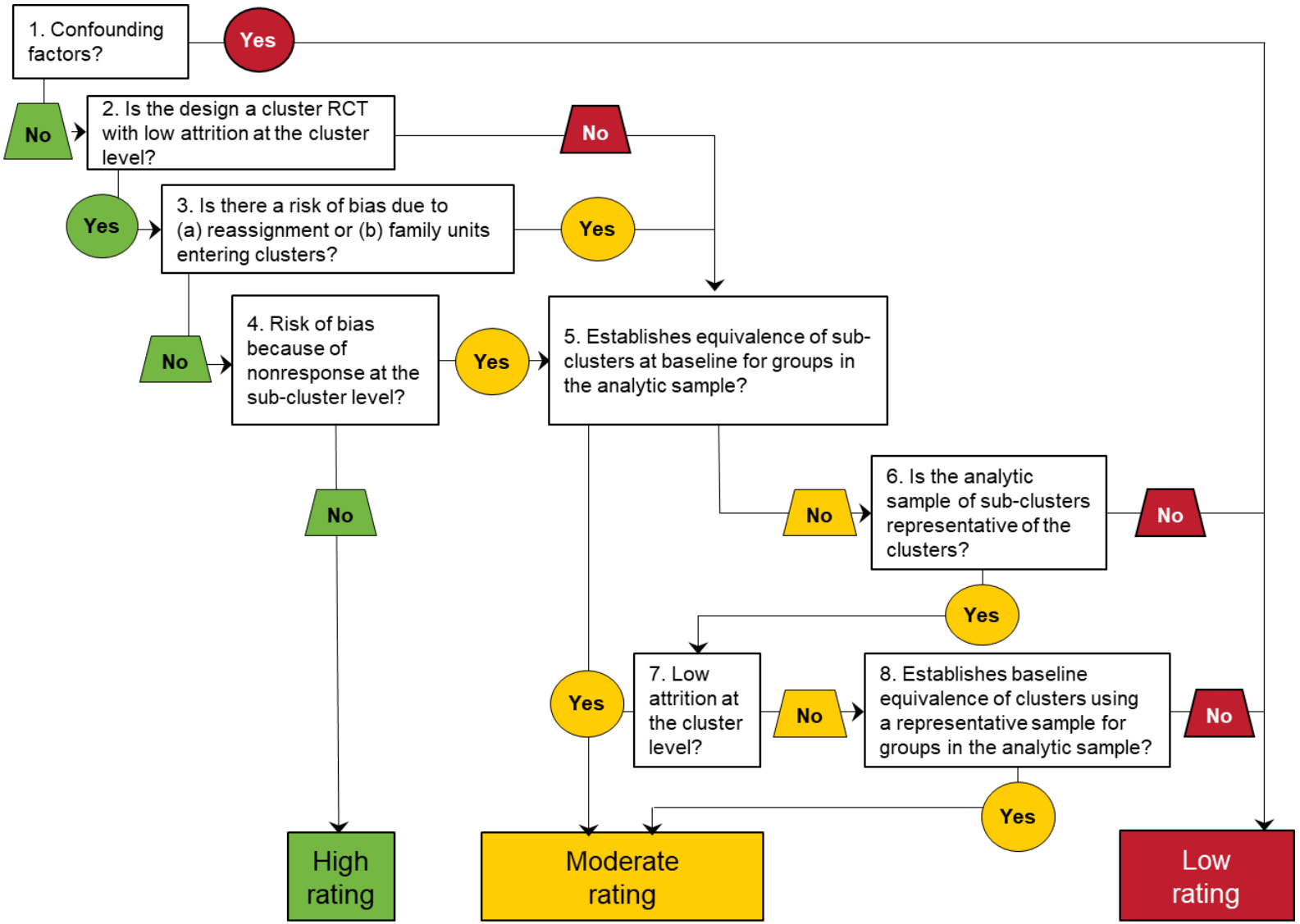
In alignment with Version 4.1 of the WWC standards and procedures (U.S. Department of Education 2020a; 2020b), HomVEE follows eight steps to review and assign ratings to cluster RCTs and NEDs (Exhibit III.14).

Step 1: Are there any confounding factors?

HomVEE uses the same approach to confounding factors within cluster RCTs and NEDs that it uses with RCTs that randomize at the family-level (See Chapter III, Section B.1 on standards for reviewing RCTs). If a confounding factor is identified—for example, the design includes only one cluster in the intervention

[49] Special reliability requirements apply to SCD research. See Appendix D.

Exhibit III.14. Steps in the review process for rating findings from cluster RCTs and NEDs



group and one cluster in the comparison group—it automatically receives a low rating and is not reviewed further.

Step 2: Is the design a cluster RCT with low attrition at the cluster level?

Attrition at the cluster level measures the loss of entire clusters from the sample that was assigned to intervention and comparison conditions (randomly in the case of RCTs). A cluster is lost when it does not contribute any outcome data to the analytic sample. HomVEE assesses attrition at the cluster level using the WWC boundary described in Chapter III, Section B.1 on standards for reviewing RCTs (Exhibits III.7 and III.8). Only cluster RCTs with low attrition at the cluster level can receive a high rating. If the design is a cluster RCT and attrition at the cluster level is low, reviewers continue to Step 3. If the design is a cluster RCT and attrition at the cluster level is high, or the design is a cluster NED, reviewers proceed to Step 5.

Step 3: Is there a risk of bias due to (a) reassignment or (b) family units entering clusters?

To receive a rating of high, cluster RCTs must not have risk of bias due to either one of two distinct events: (1) reassignment or (2) family units entering the clusters after ransom assignment. HomVEE considers that reassignment always poses a risk of bias because reassignment compromises the integrity of the random assignment. However, HomVEE does not assume that family units entering clusters after random assignment, referred to as **joiners**, always pose a risk of bias. That is because joiners do not always change the composition of the sample originally assigned to the intervention and comparison conditions in a cluster RCT, and therefore they do not always threaten the integrity of the random assignment in a cluster RCT.

Reassignment occurs when researchers move sample members (either clusters or family units) from their originally assigned condition to the other condition in the design after random assignment. An example is when, after random assignment, researchers move a cluster (such as a community) from the intervention group to the comparison group. Another example is when, after random assignment, researchers move family units from the comparison group to the intervention group so that these families can have access to a home visiting program. Whenever there is reassignment in a cluster RCT, the composition of the sample originally assigned to the intervention and comparison condition changes, which is a threat to the integrity of the random assignment in the cluster RCT.

Family units entering the clusters in an RCT after random assignment can compromise random assignment and therefore pose a risk of bias if the joiners to clusters in the intervention group are different to the joiners to the clusters in the comparison group. While joiners can differ across intervention and comparison clusters because they joined a specific cluster based on the availability of the intervention in that cluster, this is not likely to be the case in most studies HomVEE will review. For example, in a design in which researchers randomly assign communities (such as neighborhoods or counties) to an intervention group with access to a home visiting program or a comparison group with no access to the home visiting program, families moving into the study communities after random assignment might do it because of reasons such as work, availability of housing, proximity to relatives, and/or the school system in the community. It is unlikely that these families move into the communities in the study ***because*** of the availability of the specific home visiting program, even if these families know about the availability of the program in some communities. If families who move into the intervention communities do not move based on the availability of the program, they will not be systematically different from those who move into the comparison communities and they will not pose a risk of bias. Because moving into a new

community is a costly activity that is unlikely to be influenced by access to a home visiting intervention, HomVEE expects that joiners in most cluster RCTs it reviews will not pose a risk of bias and will presume that any joiners to clusters in the intervention group do not differ systematically from the joiners to the clusters in the comparison group.

However, HomVEE may sometimes have reason to suspect that families in some cluster RCT do move into a specific community because of the availability of the home visiting program and joiners in the intervention group can systematically differ from the joiners in the comparison group. For example, the home visiting intervention may be highly attractive or the study may be designed in such a way that it is not very costly to join a cluster. In such a case, joiners in the intervention group could be families that are highly motivated to obtain services from home visiting programs while the joiners in the comparison group might not be. This would make the joiners differ across the intervention and comparison groups and therefore joiners will pose a risk of bias when included in the analytic sample. .

Even in studies where HomVEE determines that joiners might pose a risk of bias, additions to the family units are never considered to be joiners, and HomVEE does not consider these additions to pose a risk of bias:

- A mother in the sample gives birth to more than one child (twins or triplets, for example)
- A new baby is born into a family unit in the sample and the home visiting model specifies that all the children in a household must be involved in the assessment of outcomes
- An adult (for example, the mother's partner, or a relative) joins the household

If randomization is not compromised by reassignment, and there are no joiners present in the analytic sample who pose a risk of bias, reviewers continue to Step 4. If randomization is compromised or joiners are present who pose a risk of bias, reviewers proceed to Step 5.

Step 4: Is there a risk of bias because of nonresponse at the sub-cluster level?

Nonresponse at the sub-cluster level (or family-unit level) in cluster RCTs refers to the difference between the sub-clusters present in a reference sample, and the sub-clusters present in the analytic sample at the time the outcome is assessed. The sub-clusters present in the analytic sample are those who contribute data to the outcome measure. The reference sample—the benchmark sample to include in the denominator of the calculation of sub-cluster nonresponse—can differ depending on the risk of bias associated with joiners.

If HomVEE reviewers determine that joiners do not pose a risk of bias, then the reference sample is the sample of sub-clusters present in nonattriting clusters at the time of random assignment. If HomVEE determines that joiners do not pose a risk of bias, then the reference sample can be either one of the following two samples:[50]

1. The sub-clusters present in nonattriting clusters at the time of random assignment, or,
2. The sub-clusters in nonattriting clusters at follow-up.

[50] This aligns with the WWC Standards Handbook, Version 4.1, Section II.B, Randomized controlled trials and quasi-experimental designs: Cluster-level assignment (U.S. Department of Education 2012b).

If the reference sample is the original randomized sample, this step measures sub-cluster level attrition. But because the reference sample can differ from the randomized sample, this step is described as measuring sub-cluster nonresponse, rather than sub-cluster attrition.

HomVEE assesses the level of sub-cluster nonresponse using the WWC boundary described in Chapter III, Section B.1 on HomVEE's standards for reviewing RCTs (Exhibits III.7 and III.8). In further alignment with Version 4.1 of the WWC Standards, HomVEE measures sub-cluster level nonresponse within the sample of nonattriting clusters. That is, the sub-clusters in clusters that are not in the analytic sample are not included in the calculation of sub-cluster level nonresponse.

If there are no confounds, attrition at the cluster level is low, randomization is not compromised, and there is no risk of bias because of nonresponse at the sub-cluster level, findings can receive a rating of high. If there are no confounds, but attrition at the cluster level is high, or randomization is compromised, or there is risk of bias because of nonresponse at the sub-cluster level, the review proceeds to Step 5. The findings can still receive a rating of moderate if they satisfy the requirements described in Steps 5 through 8.

Step 5: Is equivalence of sub-clusters at baseline established for groups in the analytic sample?

If the analyses do not include imputed data, cluster NEDs and cluster RCTs with no confounds but with high attrition at the cluster level, compromised randomization, or high nonresponse at the sub-cluster level, can receive a rating of moderate if they satisfy HomVEE's baseline equivalence requirement, as described in Section B.2 of this chapter.[51] Baseline equivalence of sub-clusters must be established on race/ethnicity, socioeconomic status, and outcomes assessable at baseline.

Steps 6 through 8: Additional requirements on the analytic sample of sub-clusters and clusters

Cluster NEDs (or, cluster RCTs that have a high risk of bias because of sample loss or compromised randomization) that do not satisfy the baseline requirement for sub-clusters in groups in the analytic sample, can still receive a moderate rating if they meet additional requirements for the analytic sample of sub-clusters and clusters. These additional requirements are necessary to minimize the risk of bias because it is possible that findings represent a combination of (1) the effect of the intervention on sub-clusters and (2) a composition effect caused by different types of sub-clusters entering the intervention and comparison clusters after random assignment. To meet the additional requirements, cluster NEDs and cluster RCTs with high risk of bias because of sample loss or compromised randomization must analyze sub-clusters that are representative of the clusters in the analytic sample. In addition, these designs must also either have low cluster-level attrition or must satisfy a requirement for the baseline equivalence of clusters in the intervention and comparison groups in the analytic sample. Steps 6 through 8 describe the additional requirements in more detail.

Step 6: Is the analytic sample of sub-clusters representative of the clusters?

HomVEE assesses how representative the sub-clusters within clusters included in the analytic sample are of all sub-clusters present in the clusters at follow-up. The sub-clusters or family units in the analytic sample are not representative of the clusters if their overall response rate at follow-up is poor or if the difference in the response rates of sub-clusters in the intervention and comparison groups is high.

[51] If the analyses include imputed data, cluster RCTs must be reviewed following the review process described in Appendix E.

To assess the representativeness of clusters, reviewers first compute nonresponse among the sub-clusters in the clusters at follow-up. In this calculation, the numerator is the number of sub-clusters present in nonattriting clusters *at follow-up* that do not contribute outcome data to the analytic sample; and the denominator is the total number of sub-clusters in nonattriting clusters *at follow-up*. For example, in a design in which neighborhoods are randomly assigned to intervention and comparison conditions, this nonresponse calculation will use (1) the number of family units (the sub-clusters) present in nonattriting neighborhoods (the clusters) *at follow-up* that do not contribute outcome data to the analytic sample, as the numerator, and, (2) the total number of family units in nonattriting neighborhoods) *at follow-up*.

Reviewers then assess whether the calculated nonresponse is high, making the sub-clusters in the analytic sample unrepresentative of the clusters. To do this, reviewers use the attrition boundary described in Chapter III, Section B.1 on HomVEE's standards for reviewing RCTs (Exhibits III.7 and III.8). If nonresponse is low in this representative assessment, the review proceeds to Step 7. If nonresponse is high, the findings from the cluster RCT or cluster NED receive a low rating.

Step 7: Is attrition at the cluster level low?

This is the same assessment done for attrition at the cluster level from Step 2. This step is repeated because it is possible for both RCTs with low attrition at the cluster level and RCTs with high attrition at the cluster level to arrive at Step 7.

If attrition is low at the cluster level, then the findings from the RCT receive a moderate rating. If attrition is high at the cluster level, then the review proceeds to Step 8.

Step 8: Is baseline equivalence of clusters established using a representative sample for groups in the analytic sample?

To receive a moderate rating, cluster RCTs with high cluster-level attrition and cluster NEDs that do not satisfy the baseline equivalence requirement for sub-clusters must satisfy the baseline equivalence requirement, as described in Chapter III, Section B.2, for the analytic sample of clusters in the intervention and comparison group. The analytic sample of clusters consists of the clusters represented in the sample that is used to estimate findings. Then, these baseline equivalence calculations are based on the sub-clusters (1) that are within the clusters represented in the sample that is used to estimate findings and (2) that contribute baseline data.

In alignment with Version 4.1 of the WWC Standards, the characteristics on which HomVEE will assess baseline equivalence of clusters may differ from those used to assess baseline equivalence of sub-clusters. Examples of such characteristics include race/ethnicity, socioeconomic status, and outcomes that are measured at the cluster level. In addition, HomVEE will consult with subject matter experts and ACF about whether sub-clusters contributing baseline data to assess baseline equivalence of clusters must be the same sub-clusters contributing outcome data to the analysis. Specifically, HomVEE will consult with experts and ACF about the following:

- **Whether the baseline equivalence requirement can be met using data from an earlier assessment of the same cohort of sub-clusters in the analytic sample within the same clusters.** For example, for designs that assigned neighborhoods to conditions, HomVEE will consult with experts and ACF on whether the baseline equivalence requirement could be satisfied for an analytic sample of family units in 2019 (for example, those that enrolled in the study in 2019 and completed the baseline assessment in 2019) using data from the previous year (2018) on that same cohort of

family units. Some characteristics of family units could change over time (for example, socioeconomic characteristics such as income level, earnings, and educational attainment), so the data on those characteristics in 2018 and 2019 might not necessarily be the same for the same group of family units.

- **Whether the baseline equivalence requirement can be met using data from an earlier cohort of sub-clusters within the same clusters.** For example, for neighborhood-level assignment studies, HomVEE will consult with experts and ACF on whether the baseline equivalence requirement could be satisfied for an analytic sample of family units that enrolled in the study in 2019 using another cohort of family units that experienced a birth and lived within the same neighborhoods in 2018.
- **The maximum elapsed time that is allowed between the collection of baseline and outcome data.** As more time elapses between the collection of baseline and outcome data, the relevance of the baseline data may become weaker. For example, if outcomes are measured in 2019 for family units but the available baseline data were collected for the same group of family units a few years earlier, for example, in 2015, there may be less overlap in the 2019 and 2015 samples than if we could use data from a more adjacent period (for example, if the baseline data were collected in 2018).

Independently of the level of analysis, the baseline equivalence requirement for clusters can be satisfied with means and standard deviations at the sub-cluster and cluster levels, in any combination, as long as the weighting of the means is consistent with the weighting used in the analysis. Whenever possible, HomVEE will use standard deviations at the sub-cluster level. Any required statistical adjustments must be made using data at the same level as those used to assess baseline equivalence.

To meet the baseline equivalence requirement for the analytic sample of clusters, the sub-clusters with baseline data must also be representative of the clusters contributing to the impact analysis. This is assessed by dividing the number of sub-clusters contributing baseline data by the number of sub-clusters in the clusters at the time of the baseline equivalence assessment, and then comparing this calculation with the boundary described in Chapter III, Section B.1 on HomVEE's standards for reviewing RCTs (Exhibits III.7 and III.8). If representativeness is high, and baseline equivalence of the analytic sample of clusters is established, findings from the cluster RCT receive a rating of moderate. Otherwise, they receive a rating of low.

Cluster correction

If the unit of assignment is different from the unit of analysis, the analysis must account for this clustering (for example, if ZIP codes are assigned to home visiting or comparison conditions, but family-level data are analyzed, not ZIP code aggregated data). Without such a correction, the statistical significance of the findings may be overstated. That is, a finding could be misclassified as statistically significant, but it might not be statistically significant 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, based on a summary of behavioral and attitudinal outcomes (WWC Procedures Handbook, Version 4.1, Section VI.A.2: Clustering correction for "mismatched" analyses). If HomVEE does not have enough information to make the correction, the uncorrected findings will be excluded from the review.

2. Repeated measures analyses

In **repeated measures analyses** (Exhibit III.15), researchers measure the research sample at several time points to chart its growth over the course of the intervention and, sometimes, beyond. HomVEE reviews

repeated measures analyses in manuscripts about RCT and NED studies when they satisfy certain eligibility requirements, including the availability of findings for individual time points.[52]

HomVEE will consider analytic approaches to be repeated measures analyses in manuscripts about RCT and NED studies whenever impacts are measured at multiple (two or more) time points or whenever the analysis method is one of those named in Exhibit III.15. If it is not possible to determine from the information provided in a manuscript whether a particular analytic approach should be reviewed as a repeated measures analysis, HomVEE will consult with subject matter experts and ACF to determine how to proceed with the review.

Exhibit III.15. Examples of repeated measures analyses HomVEE will not review unless authors provide results for each time point

These are examples and not an exhaustive list. Authors may use different names for the same method. For example, "growth curve analysis" might sometimes mean "multilevel linear modeling."

- Growth curve analyses
- Multilevel or hierarchical linear modeling (with observations over time nested within individuals)
- Repeated measures ANOVA or ANCOVA
- Latent growth curve models
- Generalized linear mixed models

Reviewers will take the steps described below to review repeated measures estimates and report the results of those reviews. Reviewing impact findings at each point in time included in repeated measures analyses enables HomVEE to assess attrition at each point in time for manuscripts about an RCT, and to define the analytic sample at each time point and establish the equivalence of characteristics between groups for manuscripts about RCTs with high attrition or about NEDs. Further, this approach enables HomVEE reviewers to identify potential threats to the research design at each point in time, including confounding factors. And it provides a consistent and fair approach to assessing statistical significance across studies.

a. Step 1: Review each finding and assess risk of bias at each time point.

HomVEE only reviews findings from repeated measures analyses with multiple follow-ups when the findings are available for individual time points. Some repeated measures analyses combine outcome measures from multiple time points into a single impact estimate. For example, a repeated measures ANOVA might combine scores across 3-month, 6-month, and 9-month follow-up periods into one impact estimate.[53] When findings for each follow-up are not provided in the manuscript, HomVEE will contact authors to request those findings. In the example above, this would mean requesting findings at each of

[52] Researchers might think of two other types of designs as repeated measures approaches. First are multiple baseline designs, which HomVEE reviews using the single-case design standards (see Appendix D). Second are interrupted time series designs, with or without comparison groups; HomVEE does not review research with an interrupted time series design.

[53] A repeated measures ANOVA or ANCOVA that just has one baseline and one follow-up measure of the outcome (that is, a difference in differences analysis) is eligible for review. If the outcome is rated high or moderate, reviewers report the time*treatment effect, which measures whether the average change in the outcome from the pre- to post-follow-up points differs in the two groups.

the 3-, 6-, and 9-month time points. If the only finding available across the study and the author query is a combined impact estimate, that impact estimate is not eligible for review.

HomVEE reviews findings for each of the time points in the repeated measures approach separately (including an assessment of attrition and baseline equivalence for each outcome at each time point). In addition, HomVEE reviews impacts at each time point relative to baseline (not relative to other follow-up points). This means reviewing findings from each time point separately, even if the time points overlap. For example, if a manuscript reports the same outcome measured at both 6 and 12 months after the home visiting services were provided, HomVEE reports the findings for each time point separately even though the baseline-to-12-month follow-up overlaps with the baseline-to-6-month follow-up. This approach is consistent with HomVEE's approach to reviewing other group-design research in RCT and NED studies, in which a research team might follow a sample and report findings for different follow-ups across different manuscripts.

Step 2 describes how ratings are assigned to findings from follow-ups rated moderate or high.

b. ***Step 2: Assign ratings and identify which findings are eligible for reporting.***

HomVEE assigns ratings of high, moderate, or low to each finding at each of the time points (follow-ups) in the repeated measures approach, following the processes for assigning ratings to findings from RCTs and NEDs (Chapter III, Section B on HomVEE's standards for reviewing eligible designs and outcomes). If all findings rate low at all time points, the manuscript rates low.

HomVEE reports high- or moderate-rated findings for each time point or follow-up under the three following scenarios, which are presented in the order of HomVEE's preference to minimize additional requests to authors (including re-analysis requests). The three scenarios are:

- **Scenario 1 (preferred): HomVEE reports author-reported findings at each time point.** If authors report (in the manuscript under review or through an author query) findings at each time point, HomVEE reports the findings at each time point that received a rating of high or moderate. If in the manuscript under review authors provide estimates for impacts at each time point, HomVEE will report the findings that rate high or moderate for each separate time point.[54]
- **Scenario 2 (alternate): HomVEE calculates unadjusted time point findings based on details the contractor review team has from authors.** Sometimes, authors report (in the manuscript and/or in an author query) all the information necessary to rate findings at each time point separately (for example, information to determine the level of attrition and establish baseline equivalence), but they do not report impact estimates separately for each time point. In such cases, HomVEE can calculate findings—that is, unadjusted effect sizes and *p*-values—for each time point,[55,56] using author-provided means and standard deviations for each outcome at each time point. This approach is appropriate only when unadjusted time point estimates would be accepted by HomVEE's RCT or

[54] If the findings for each time point (Scenario 1) were calculated based on the trend from the growth curve analysis, HomVEE prefers to report estimates based on Scenario 2, if possible. However, if Scenario 2 does not apply, HomVEE will accept point-in-time estimates calculated based on the trend (Scenario 1) and not ask authors to conduct additional analyses (Scenario 3).

[55] Means and standard deviations can only be used to calculate valid impact estimates if the analysis in the manuscript did not need to include controls to receive a high rating.

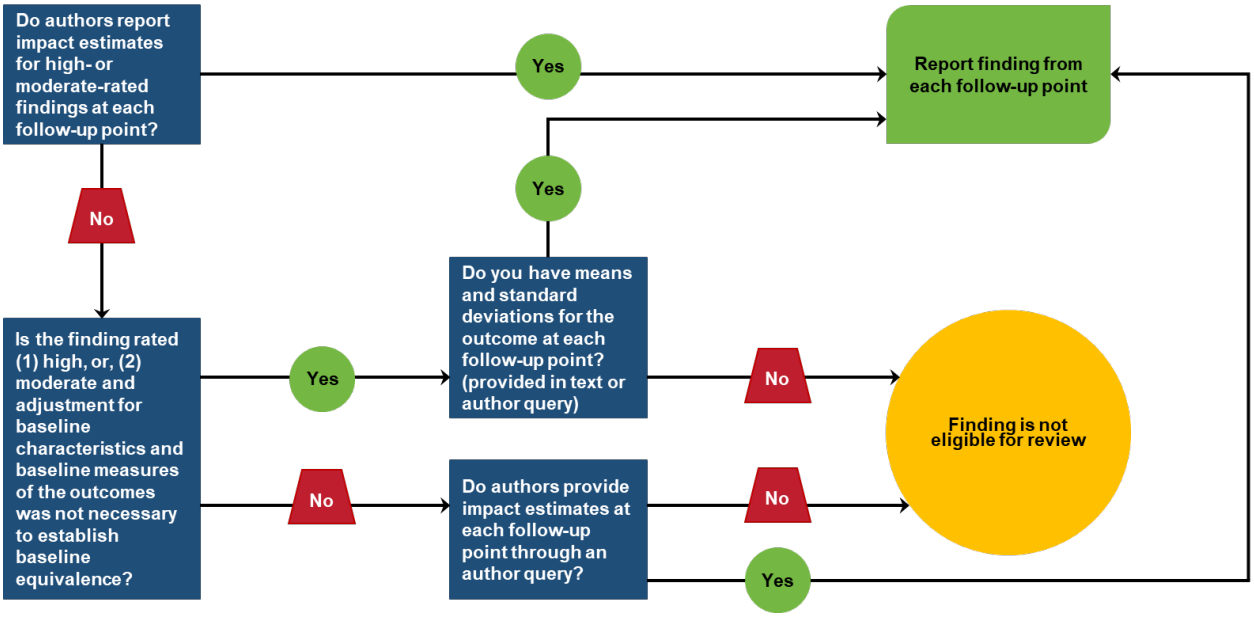
[56] Means and standard deviations can only be used to calculate valid impact estimates for a moderate-rated manuscript if the analysis in the manuscript did not need to control for a baseline outcome that was assessable at baseline.

NED standards. Specifically, this approach is appropriate if outcomes are either (1) rated high or (2) rated moderate AND adjustment for baseline characteristics and baseline measures of the outcomes was not necessary to establish baseline equivalence.[57]

- **Scenario 3 (last resort): HomVEE asks authors to calculate adjusted time point findings.** If authors do not report impact estimates separately for each time point and adjustment for baseline characteristics and baseline measures of the outcomes is necessary for a moderate rating, it is not appropriate for HomVEE to calculate and report unadjusted effect sizes and *p*-values. In these cases, HomVEE will ask authors to calculate impacts for each time point, adjusted for baseline characteristics and outcomes, and will report those author-calculated findings.[58] This requires more effort from authors than the other two scenarios do, and is therefore the least preferred option. If authors do not provide adjusted impacts for each time point, the findings will not be considered eligible for review. HomVEE will exclude from its review of a repeated measures analysis any time points for which an impact cannot be reported because neither author-provided nor HomVEE-calculated estimates are available.

Exhibit III.16 illustrates the flow of reviewer and author interactions related to reporting results from manuscripts about repeated measures studies.

Exhibit III.16. Decision flow for HomVEE reporting of high- or moderate-rated outcomes from repeated measures analyses



Note: Manuscripts about repeated measures studies for which all follow-ups rate low will be rated as low. If the findings for each time point (Scenario 1) were calculated based on the trend from the growth curve analysis, HomVEE prefers to report estimates based on Scenario 2, if possible. However, if Scenario 2 does not apply HomVEE will accept point-in-time estimates calculated based on the trend (Scenario 1) instead of making authors conduct additional analyses (Scenario 3).

[57] This includes cases in which the outcome(s) was(were) not assessable at baseline.
[58] HomVEE will provide a clear and detailed request for the information needed, but HomVEE is unable to provide technical assistance to authors about research design and analysis.

3. Structural equation models

Structural equation models (SEMs) are a statistical modeling technique that analyzes the structural relationships between variables, often including both observed and unobserved, or latent constructs. SEMs typically combine several statistical techniques such as path analysis and factor analysis (Hox and Bechger 1998). However, the term "SEMs" is also used to describe models that only include observed constructs in a path analysis or models that only create latent variables through a factor analysis. In other words, path analysis and factor analysis can be thought of as special cases of SEM (Hox and Bechger 1998). Researchers often use SEMs to estimate both the magnitude and significance of causal connections between variables. In addition to baseline covariates that are typically incorporated into regression models, SEMs often include multiple outcomes (sometimes from different follow-up periods).

HomVEE reviewers apply the standards below when authors describe their analysis using any of the following terms commonly used to denote SEMs **and** when the authors' analysis is accompanied by a path diagram (see Exhibit III.17):

- Path/pathway analysis
- Moderating analysis
- Mediating analysis
- Factor analysis (refers to the measurement portion of an SEM)
- Latent growth modeling
- LISREL (a software package that supports SEM analyses)
- Simultaneous equation model
- Structural equation model
- Analyses that estimate so-called indirect effects

a. Confirm that a diagram accompanies the SEM analysis

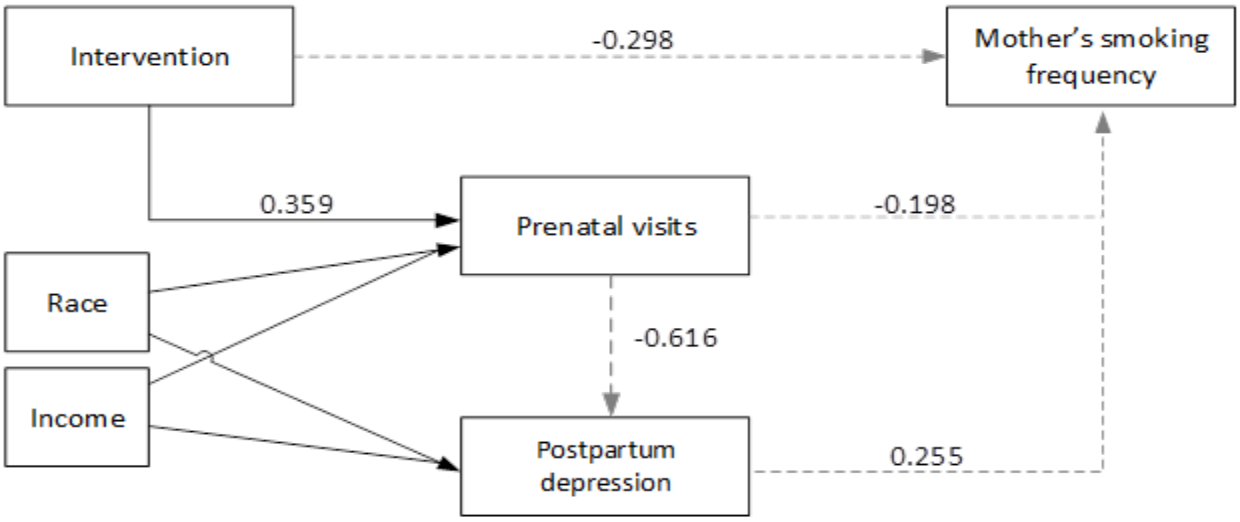
Only identified SEMs that are accompanied by a path diagram are eligible for review by HomVEE.

Because HomVEE relies on diagrams to identify effects that are eligible for review (see next section), the manuscript must present a diagram that summarizes the relationships between the intervention and other variables. It is not enough, for HomVEE review purposes, for a manuscript to describe the SEMs through equations. However, if authors report regression equations, there should be an exact correspondence between the regression equations and the diagram. See Exhibit III.17 for an example of an SEM diagram. It depicts the relationship between the intervention, other factors, and the outcomes (the center, bottom, and right boxes) by illustrating the effects (the paths, shown as arrows in this diagram) of the intervention and other factors on the outcomes. Only some of these effects are findings that are eligible for review by HomVEE, as discussed in the next section. If the manuscript under review does not include a path diagram and reviewers determine that the authors have applied a SEM analysis approach, HomVEE will request one in an author query. If authors do not provide a path diagram, the review of that manuscript stops.

HomVEE also checks whether authors have reported essential information: degrees of freedom, number of free parameters (those being estimated), and whether any parameters were fixed. Reviewers then check whether the model is **identified:** that is, the degrees of freedom are greater than the parameters to be

estimated. If HomVEE finds that an SEM is not identified, findings are not considered eligible for review, and the review of that manuscript stops.

Exhibit III.17. Depiction of structural equation model outcomes that would be eligible for review by HomVEE



Note: Only the solid arrow connecting the intervention to prenatal visits would be a finding eligible for review by HomVEE. Dashed arrows represent findings that would be ineligible for review by HomVEE.

b. Identify effects in the diagram that are eligible for HomVEE to review

When reviewing an SEM diagram to identify which outcomes to review, HomVEE reviewers ask: In the SEM diagram, is there a direct pathway from the intervention to the outcome? AND: Are there no pathways leading *to* that outcome *from* another outcome? If both answers are yes, that outcome is eligible for review. Outcomes for which reviewers can answer these two questions affirmatively are eligible even if they have pathways pointing toward them from baseline characteristics (such as race or SES), as in the case of prenatal visits.

Exhibit III.17 identifies outcomes that are eligible and ineligible for review by HomVEE based on these criteria.[59] HomVEE does not have standards for mediated, indirect or total (if they consist of direct plus indirect/mediated) effects examined within these models, and thus does not rate or review them. However, if authors describe their approach as a mediating analysis, it can still be eligible for review if a path diagram is provided and it allows reviewers to verify that the analyses estimated direct paths from treatment to outcomes and that those direct paths are not mediated by another outcome. If a path diagram is not included in the manuscript, HomVEE will request one from the author.

For example, as illustrated earlier in Exhibit III.17, the effect of the home visiting intervention on prenatal visits is eligible for review by HomVEE because it has only one pathway leading to it (from the home visiting intervention). In this example, the effect of the home visiting intervention on postpartum

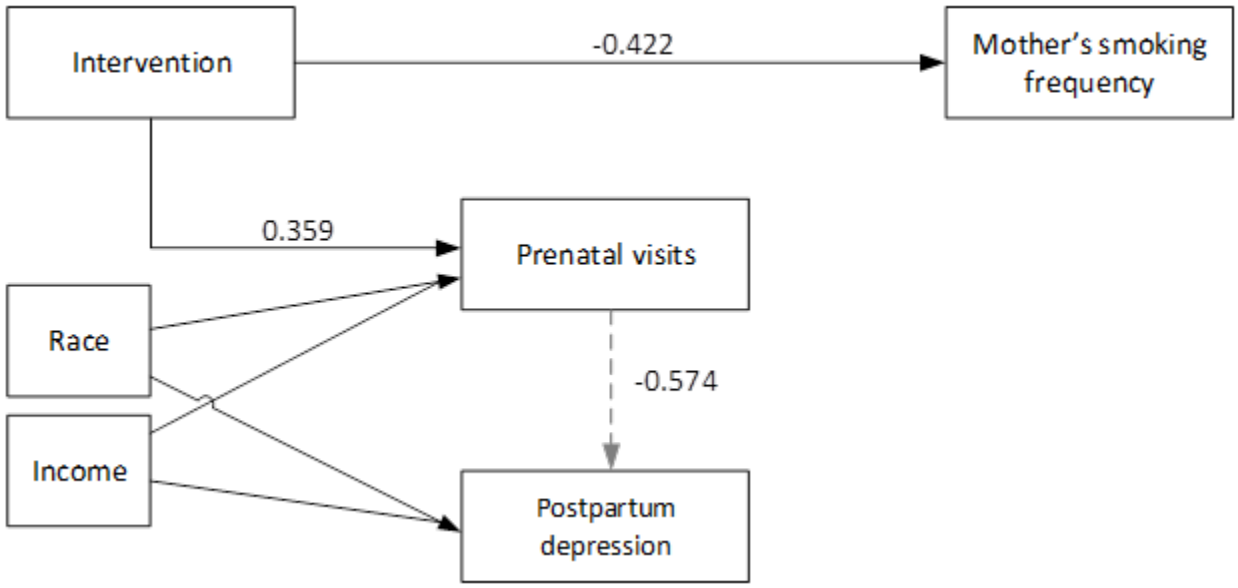
[59] Reviewers also check whether the outcome is otherwise eligible for review by HomVEE, such as whether it falls into one of the eight domains HomVEE examines. See Chapter III, Section A.4 for more information. HomVEE recognizes that the other paths in an SEM analysis, which show relationships between outcomes, may be of interest to researchers and the field. But those relationships do not answer the core question HomVEE examines, that of whether the home visiting intervention itself causes changes in an outcome of interest.

depression will not be reviewed, because it is not connected to the intervention except through prenatal visits. In other words, prenatal visits mediate the effect of the intervention on postpartum depression. Although the total effect of intervention on postpartum depression can be calculated based on the information in Exhibit III.17 (as 0.359 multiplied by -0.616), the HomVEE review cannot calculate the statistical significance of this total effect of the home visiting intervention on postpartum depression without additional information about the covariance. HomVEE requires a measure of statistical significance for all findings to be able to apply the HHS criteria to identify evidence-based models.

Mother's smoking frequency has a pathway linked directly to the intervention, but prenatal visits and postpartum depression also link to it. Therefore, the effect of intervention on the smoking frequency is not independent from the effects of intervention on two other outcomes (prenatal visits and postpartum depression). For this reason, the effect of intervention on mother's smoking frequency is not eligible for review.

Note that when the prenatal visits and postpartum depression outcomes would not have pathways to mother's smoking frequency, then both the pathway linking intervention and mother's smoking frequency *and* the pathway linking intervention to the prenatal visits would be eligible for review (Exhibit III.18). Both these effects would satisfy the eligibility criteria as they are *direct and* they do not have pathways leading *to* them *from* other outcomes.

Exhibit III.18. Example of a SEM in which two outcomes are eligible for review by HomVEE



Note: Both solid arrows connecting the intervention to prenatal visits and the intervention to mother's smoking frequency would be findings eligible for review by HomVEE.

c. Apply HomVEE criteria to the SEM findings that were eligible for review

HomVEE reviewers apply the usual HomVEE criteria described earlier in this chapter to the SEM outcomes that are eligible for review. This includes calculating attrition, assessing baseline equivalence, examining how authors of manuscripts about SEM studies addressed missing data, and issuing any necessary queries to authors to clarify details about their research.

D. Imputation and handling of missing data

Imputation is a statistical approach authors use to estimate missing data points when data are missing from a study for some cases overall or at some follow-up points. If authors use this approach on RCTs, HomVEE still relies on the analytic sample with measured outcomes to assess attrition and baseline equivalence in the manuscript. HomVEE will review imputed *findings* and incorporate the resulting *p*-values and standard errors into the review if authors use acceptable imputation (see Exhibit III.19 for a list of the accepted methods, and Appendix E, Table E.1 for a detailed description of each method and the requirements for its application). Findings from RCTs with low attrition that use imputed data in analyses are eligible for a high rating. Findings from NEDs and RCTs with high attrition based on analyses of imputed data are eligible at best for a moderate rating, if they satisfy baseline equivalence requirements that account for the missing or imputed data. These standards derive from WWC Version 4.1 standards (see Appendix E for details on the review process for findings from designs with missing outcome or baseline data).[60]

Exhibit III.19. List of methods for addressing missing data that HomVEE accepts

- Complete case analysis
- Regression imputation (must be conducted separately by treatment status)
- Maximum likelihood (including expectation maximization and full information maximum likelihood)
- Nonresponse weights (must be conducted separately by treatment status; acceptable only for missing outcome data, not for missing baseline data)
- Replacing missing data with a constant combined with including a missing data indicator (acceptable only for missing baseline data, not for missing outcome data)

Note: Please see Appendix E for a detailed description of each method and the requirements for its application.

[60] See the WWC standards handbook at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf.

This page has been left blank for double-sided copying.

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.

Colman, Silvie. "Estimating Program Impacts for a Subgroup Defined by Post-Intervention Behavior: Why Is It a Problem? What Is the Solution?" Evaluation technical assistance brief for OAH & ACYF teenage pregnancy prevention grantees. Washington, DC: Office of Adolescent Health, U.S. Department of Health and Human Services, December 2012. Available at https://www.hhs.gov/ash/oah/sites/default/files/ash/oah/oah-initiatives/assets/estimating_programs_brief.pdf. Accessed June 24, 2020.

Duncan, G. J., and J. Brooks-Gunn (eds.). *Consequences of Growing Up Poor*. New York, NY: Russell Sage Foundation, 1997.

Eckenrode, J., B. Ganzel, C.R.J. Henderson, E. Smith, D.L. Olds, J. Powers, R. Cole, H. Kitzman, and K. Sidora. "Preventing Child Abuse and Neglect with a Program of Nurse Home Visitation: The limiting Effects of Domestic Violence." *JAMA,* vol. 284, no. 11, 2000, pp. 1385–1391. doi:10.1001/jama.284.11.1385.

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, no. 2, 2012, pp. 140–158.

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, pp. 425–445.

Gomby, D. "Home Visitation in 2005: Outcomes for Children and Parents." Sunnyvale, CA: Committee for Economic Development, Invest in Kids Working Group, 2005.

Gomby, D. S., P. L. Culross, and R. E. Behrman. "Home Visiting: Recent Program Evaluations—Analysis and Recommendations." *Future of Children,* vol. 9, no. 1, 1999, pp. 4–26.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. "Statistical Tests, *p*-Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology,* vol. 31, no. 4, 2016, pp. 337–350.

Guise, J.M., M.E. Butler, C. Chang, M. Viswanathan, T. Pigott, and P. Tugwell. "Complex Interventions Workgroup. AHRQ Series on Complex Intervention Systematic Reviews –Paper 6: PRISMA-CI Extension Statement & Checklist." *Journal of Clinical Epidemiology*. vol. 90, 2017, pp. 43–50.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis, 15*(3), 199-236.

Hox, J.J,. and T.M. Bechger. "An Introduction to Structural Equation Modeling." *Family Science Review*, vol. 11, 1998, pp. 354–373. Available at http://joophox.net/publist/semfamre.pdf. Accessed June 24, 2020.

Imai, K., G. King, & E.A. Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* vol. 171, no. 2, 2008, pp. 481-502.

Jackson, M. I., K. Kiernan, & S. McLanahan. Maternal Education, Changing Family Circumstances, and Children's Skill Development in the United States and UK. *The ANNALS of the American Academy of Political and Social Science*, vol. 674, no.(1), 2017, pp. 59-84.

Leon, D. A. "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)*, vol. 34, no. 4, August 2011, pp. 200–208.

MacKinnon, D. P. "Integrating Mediators and Moderators in Research Design." *Research on Social Work Practice*, vol. 21, no. 6, 2011, pp. 675–681.

Myers, D., W. C. Baer, and S. Y. Choi. "The Changing Problem of Overcrowded Housing." *Journal of the American Planning Association*, vol. 62, no. 1, 1996, pp. 66–84. doi:10.1080/01944369608975671.

McGowan, Jessie, Margaret Sampson, Douglas M. Salzwedel, Elise Cogo, Vicki Foerster, and Carol Lefebvre. "PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement." *Journal of Clinical Epidemiology*, vol. 75, 2016, pp. 40–46.

Moher, D., L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, and L.A. Stewart. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement." *Systematic Reviews*, vol. 4, no. 1. doi: 10.1186/2046-4053-4-1

National Home Visiting Resource Center. "Home Visiting Yearbook: An Overview." 2018. Available at https://www.nhvrc.org/wp-content/uploads/NHVRC_Yearbook-Summary_2018_FINAL.pdf. Accessed June 24, 2020.

National Home Visiting Resource Center. "Home Visiting Yearbook: An Overview." 2019. Available at https://live-nhvrc.pantheonsite.io/wp-content/uploads/NHVRC_Yearbook_Summary_2019_FINAL.pdf. Accessed September 11, 2020.

Patra, K., M. M. Greene, A. L. Patel, & P. Meier. Maternal Education Level Predicts Cognitive, Language, and Motor Outcome in Preterm Infants in the Second Year of Life. *American Journal of Perinatology. vol. 33, no.* 8, 2016, pp. 738–744.

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

Schochet, O. N., A. D. Johnson, & R. M. Ryan. The Relationship Between Increases in Low-income Mothers' Education and Children's Early Outcomes: Variation by Developmental Stage and Domain. *Children and Youth Services Review. vol. 109*, 2020, p. 104705.

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

Stoltzfus, E., and K. Lynch. "Home Visitation for Families with Young Children." No. R40905. Washington, DC: Congressional Research Service, 2009.

Thompson, Matthew, Arpita Tiwari, Rongwei Fu, Esther Moe, and David I. Buckley. "A Framework to Facilitate the Use of Systematic Reviews and Meta-Analyses in the Design of Primary Research Studies." Rockville, MD: Agency for Healthcare Research and Quality, 2012.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. *What Works Clearinghouse Procedures and Standards Handbook: Version 3.0*. 2013. Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf Accessed June 24, 2020.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. *What Works Clearinghouse Procedures Handbook: Version 4.1*. 2020a. Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf. Accessed June 24, 2020.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. *What Works Clearinghouse Standards Handbook: Version 4.1*. 2020b. Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf. Accessed June 24, 2020.

U.S. Department of Health and Human Services, Administration for Children and Families, Health Resources and Services Administration, Maternal and Child Health. The Maternal, Infant, and Early Childhood Home Visiting Program: Partnering with Patents to Help Children Succeed. 2019. Available at https://mchb.hrsa.gov/sites/default/files/mchb/MaternalChildHealthInitiatives/HomeVisiting/pdf/programbrief.pdf. Accessed June 24, 2020.

Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA's Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician,* vol. 70, no. 2, 2016, pp. 129–133.

World Bank. *Word Bank Indicators*. Historical classification for 2009. 2020. Available at https://datahelpdesk.worldbank.org/knowledgebase/articles/906519. Accessed on April 8, 2020.

This page has been left blank for double-sided copying.

Glossary of Terms

A

Analytic sample. The families represented by the outcome data used in the analysis, which may be different from the sample assigned to the intervention and comparison conditions at the beginning of the study.

Analytic subgroup. A subset of the sample examined in a study.

Attrition. This happens when outcome data are missing for some members of the intervention or comparison groups. Attrition can occur because sample members do not respond to surveys or are missing from administrative data sets, or it can occur for some other reason.

B

Baseline. The time before the intervention being studied is provided to the intervention group

C

Cluster randomized controlled trial (RCT). HomVEE considers designs to be cluster RCTs or non-experimental comparison group designs whenever family units are assigned (randomly in the case of RCTs) to the intervention or comparison conditions as groups (or clusters), such as a neighborhood, ZIP code, or county.

Comparison group. A group with characteristics similar to those of intervention 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 intervention group if they had not received the services from the model of interest. The more similar the characteristics of a comparison and intervention group are, the more likely it is that any difference in outcomes between the two groups can be attributed to the intervention.

Confounding factor. Any observed factor that is completely aligned with either the intervention or comparison group. This means that the factor is present in only the intervention group or the comparison group, but not both. See Appendix D for a specialized definition of confounding that HomVEE applies to SCD research.

Cox index. This index calculates an effect size for proportions. HomVEE uses it for findings from dichotomous variables, which have only two possible values (usually, 0 and 1).

D

Differential attrition. Differential attrition refers to the difference in the rate of attrition between the intervention and comparison groups.

Distinct (or nonoverlapping) samples. Two or more studies in which there are no sample members in common. For example, one study conducted from 2002 through 2004 and another conducted from 2008 through 2010.

E

Effect size. A measure of the magnitude of the difference between the intervention group and the comparison group. The effect size shows the size of the impact (or the difference between the intervention 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 by 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 intervention group (which did receive services). A positive value indicates that the outcomes for the intervention group were greater than those for the comparison group. Values of 0 indicate there is no difference, on average, between the intervention and comparison groups.

Endogenous characteristics. Characteristics of study participants that are defined by behavior that emerges after they learn whether they will be in the intervention group or the comparison group, or could theoretically be affected by a home visiting model. Therefore, there is a relationship between the assignment to the intervention and the endogenous characteristics (that is, they are not independent of each other).

Equivalence. This means that the intervention and comparison groups are similar on specified characteristics. (HomVEE recognizes that some characteristics and outcomes are not always age- or developmentally appropriate to collect.)

Evaluation. An evaluation is an analysis of a distinct implementation of an intervention. HomVEE reviews and rates evaluations that examine the effectiveness of a home visiting intervention (see Chapter III for details).

Evidence-based model. For the purposes of the HomVEE review, this handbook uses the term "evidence-based model" to refer specifically to a model that meets HHS criteria developed based on statutory requirements in the authorizing legislation for the MIECHV Program. HomVEE recognizes that other systematic reviews may use different criteria to evaluate evidence of effectiveness. Thus, an evidence-based model in the context of HomVEE might or might not meet requirements for evidence of effectiveness according to other systematic reviews.

F

Favorable finding. A finding showing a statistically significant impact on an outcome measure in a **direction** that is beneficial for children and parents. An impact could be statistically positive or negative. It is determined to be "favorable" based on the result. For example, a favorable impact could be an increase in children's vocabulary or in daily reading to children by parents, or a reduction in child maltreatment or maternal depression.

Findings. Findings summarize the effect of a home visiting model on a specific sample or subgroup, on a specific eligible outcome measure (see Chapter III), at a specific time point, from a specific analysis. A manuscript typically includes multiple findings. HomVEE rates findings (see Chapter III) and sorts manuscripts according to the highest-rated finding in the manuscript (see Chapter II).

H

Hedges' *g*. This is the ratio between the estimated impact of the intervention (the difference between the intervention and comparison group scores) and the standard deviation (the variation in scores) pooled across the intervention and comparison groups.

High rating. A rating applied to a manuscript when there is strong evidence that at least one finding reported in the manuscript is attributable to the intervention that was examined.

I

Identical samples. Two or more manuscripts that report results from an analytic sample whose entire group of participants consists of the same sample members. For example, this could be two manuscripts on the same intervention and comparison group that report findings on different outcomes.

Identified. An analysis model is identified if the degrees of freedom are greater than the parameters to be estimated.

Model **implementation summary reports.** These reports are implementation profiles about the implementation of each prioritized model. They include an overview of the model and information about prerequisites for implementation, materials and forms, estimated costs, and model contact information. Model developers or national model offices are invited to review and comment on the profiles before their release. (For evidence-based models reviewed prior to 2021, the reports also include a page describing implementation experiences in the research reviewed. These included the characteristics of participating families, location and setting, staffing and supervision, model components, model adaptations or enhancements, dosage, and lessons learned).

Imputation. Imputation is a statistical approach that authors use to estimate missing data points when data are missing from a study for some cases overall or at some follow-up points.

Intent to treat (ITT). ITT is the effect of being offered the home visiting intervention.

Intervention. In the context of HomVEE, an intervention is generally a home visiting model.

Intervention group. The sample members who receive the intervention of interest (the home visiting model).

J

Joiners. Participants (in home visiting research, usually family units) entering clusters in a cluster RCT after random assignment.

L

Low rating. A rating applied to a manuscript when there is little evidence that the reported finding is attributable, partly or as a whole, to the intervention that was examined. Manuscripts that contain no moderate- or high-rated findings receive a low rating. Otherwise, the manuscript receives the highest rating of any finding within it, and HomVEE footnotes the reason why other findings within that manuscript received a low rating.

M

Manuscript. Manuscripts describe study results. They may be published or unpublished research, such as journal articles, book chapters, or working papers. A single study may produce one or many manuscripts. Typically, one manuscript reports on only one study, although in rare cases one manuscript may include several studies, if it describes evaluations of multiple interventions or the same intervention evaluated in multiple distinct (non-overlapping) samples.

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

Mediating analyses. Mediating analyses investigate the process by which the home visiting intervention achieves its effects. These seek to answer the question: How (or under what circumstances) does the model work? Researchers examine these questions by conducting a path analysis, estimating certain types of multiple linear regression models, or running a structural equation model.

Model. HomVEE defines an **early childhood home visiting model** as an intervention in which trained home visitors meet with expectant parents or families with young children to deliver a specified set of services through a specified set of interactions. These programs are voluntary interventions that are either designed or adapted and tested for delivery in the home. During the visits, home visitors aim to build strong, positive relationships with families to improve child and family outcomes. Services may be delivered on a schedule that is defined or can be tailored to meet family needs. A model has a set of fidelity standards that describe how the model is to be implemented.

Models reviewed by HomVEE must serve pregnant women or families with children from birth to kindergarten entry (that is, through age 5), and the primary service delivery strategy must be home visiting. In addition, the model must have research that examines its effects in at least one of eight 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. (Note: These domains are inclusive of the benchmark domains and individual outcomes listed in the staututе that authorized the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) Program (Social Security Act, Section 511 [42 U.S.C. 711]).)

Moderate rating. A rating applied to a manuscript when there is some evidence that at least one finding reported in the manuscript is attributable, at least partly, to the intervention that was examined. However, other factors not accounted for in the study might also have contributed to the finding.

Moderating analyses. These are analyses that investigate the ways that specific variables influence the effectiveness of the home visiting intervention. Moderating analyses answer two questions: (1) How (or under what circumstances) does the model work? (2) Does it work equally well for different groups?

N

New information. This may discuss a study's methods or procedures.

New research. This is different from new information. It could be additional findings, new analyses of research in a previously reviewed manuscript, or an entirely new set of findings.

No effect impact. A finding that is not statistically significant.

Noncompliers are study participants who receive services they were not supposed to receive (for example, participants who were randomly assigned to the comparison group receiving the services from the home visiting model offered to the intervention group).

Non-experimental comparison group design (NED). This design uses a nonrandom process to assign sample members to an intervention group and a comparison group. Sample members can be assigned through statistical techniques that are designed to match sample members in each group, so each group has similar measurable characteristics on average; or they can be assigned based on convenience, by assigning people to groups because they are nearby or available or otherwise convenient to include.

O

Outcome domain. A group of related outcomes that measure the same or similar constructs. The HomVEE review includes eight outcome domains: (1) child health; (2) child development and school readiness; (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.

Overall attrition. Overall attrition is the combined loss of data for any sample member from either the intervention or comparison group.

Overlapping samples. Two or more manuscripts that report results from an analytic sample in which the intervention groups or comparison groups have at least some sample members in common. For example, researchers following participants over time who lose some participants in follow-ups would have a sample for the later follow-up that overlaps with the sample from the earlier follow-up.

P

***p*-value.** The probability that a difference in means (effect) at least as large as the one observed would occur by chance (when there is not a real relationship in the population). For example, a sample may show a positive mean difference, suggesting that the intervention 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 a result at least as large as the positive finding for the intervention group would be obtained by chance if, in fact, there is no true effect in the population.

Prioritization process. HomVEE's process for selecting models to review each year, which reflects the systematic review's emphasis on reviewing well-designed impact studies, examining outcomes of interest to HHS, and aligning to Maternal Infant Early Childhood Home Visiting Program criteria. HomVEE aims to identify new evidence-based models (including among previously reviewed models that HomVEE has not found to be evidence based)while continuing to update reports on models that it has already reviewed and that already are evidence based, to ensure that reported findings are up to date to the extent possible.

R

Randomized controlled trial (RCT). A study design in which sample members (children, parents, or families) are assigned to the intervention and comparison groups at random. Sample members can be assigned as individuals or as groups, depending on the study.

Reassignment. This occurs in an RCT when the researcher moves a sample member from the intervention group to the comparison group after random assignment.

Regression discontinuity design (RDD). In this design, study participants are assigned to the intervention and comparison groups using a criterion as a cutoff point. Researchers then compare participants who are some set distance above and below the cutoff point. The effect of the intervention is estimated as the difference in mean outcomes between intervention 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.

Repeated measures analyses. In these analyses, researchers measure the research sample at several time points to chart its growth over the course of the intervention and, sometimes, beyond. Authors may use different names for the method. Common examples include growth curve analyses, multilevel or hierarchical linear modeling (with observations over time nested within individuals), repeated measures ANOVA or ANCOVA, latent growth curve models, generalized linear mixed models, and averaging across time points. These are examples and not an exhaustive list.

Replicable subgroup. A subgroup defined by a characteristic that a different study could replicate with a non-overlapping sample.

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 manuscripts based on different samples.

Replicated subgroup. A subgroup that has an identical definition in two non-overlapping research samples.

S

Sample. A sample encompasses both the entire intervention group and the entire comparison group of participants included in a study. (Note: in studies that use a single-case design, the sample participants receive both the intervention and the comparison condition.)

Single-case design (SCD). In this design home visiting evaluations assign the intervention and comparison conditions to a single family or a small group of families **during certain time periods**. In single-case designs, researchers follow each study family or small group of families across several points in time. By using each individual or small group of individuals as their own comparison group, single-case designs ensure that the intervention and comparison groups have the same measured and unmeasured characteristics.

Standard deviation. A measure of the spread or variation of values in the sample. The standard deviation approximates the average distance from the mean. Smaller standard deviations indicate that the values for individual sample members are close to the mean, whereas larger standard deviations indicate there is more variation in values.

Statistical control or statistical adjustment. A statistical control or statistical adjustment is a method that researchers use to include the baseline measures in a statistical model. HomVEE allows authors to use several techniques to satisfy this requirement (see Chapter III).

Statistically significant impact. HomVEE considers a finding to be a statistically significant impact if the *p*-value of a two-sided hypothesis test of whether an effect is equal to zero is less than 0.05. That is, the finding is likely to be due to the home visiting model, rather than due to chance.

Structural equation models (SEMs). SEMs are a statistical modeling technique that analyzes the structural relationship between variables, often including latent constructs.

Study. A study evaluates a distinct implementation of an intervention (that is, with a distinct sample, enrolled into the research investigation at a defined time and place, by a specific researcher or research team).

Subgroup. A subgroup is a subset of the sample examined in a study (that is, an analytic subgroup). For example, researchers may examine how a home visiting model affects teenage mothers when there are mothers with a range of ages in their study; hence, teenage mothers would be an analytic subgroup. Sometimes researchers present subgroup findings in a manuscript alongside findings for the overall sample, and sometimes researchers prepare a manuscript based exclusively on subgroup findings from a broader study. (For HomVEE, results from teenage mothers would not be considered an analytic subgroup analysis if the overall study only enrolled teenage mothers.)

Sustained. For the HomVEE review, favorable impacts on outcomes measured at least one year after program enrollment, based on the language in the MIECHV authorizing statute (Social Security Act, Section 511 [42 U.S.C. 711]).

T

Track 1 models. Those that HomVEE has not previously found to be evidence based, as well as models that HomVEE has never reviewed.

Track 2 models. Those that HomVEE has already reviewed, and that already are evidence based.

Treatment on the treated (TOT). TOT is the effect of actually receiving the home visiting intervention.

U

Unfavorable or ambiguous findings. A finding showing a statistically significant impact on an outcome measure in a direction that may indicate potential harm to children and/or parents. An impact could be statistically positive or negative. Outcomes with either arithmetic sign could be unfavorable or ambiguous. Although some outcomes are clearly unfavorable, for other outcomes it is less clear which direction is desirable. For example, an increase in children's behavior problems is clearly unfavorable. However an increase in the number of days that mothers are hospitalized after birth is ambiguous. It could be viewed as unfavorable because it indicates that mothers have more health problems, but it could also indicate that mothers have increased access to needed health care because they are participating in a home visiting program.

Unique findings. Findings that report results on a different outcome, sample or subgroup, or time period, or with a different analytic approach, than findings reported in other manuscripts about the same home visiting model.

W

Well-designed. Well-designed impact studies are those whose design and execution suggest that some or all of the findings were due to the home visiting model rather than other factors (see Chapter III).

This page has been left blank for double-sided copying.

Appendix A

PRISMA-P and PRISMA-CI Elements

This page has been left blank for double-sided copying.

The tables in this appendix address each relevant section of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis for Protocols (PRISMA-P; Moher et al. 2015) and the methods section of the PRISMA for Complex Interventions (PRISMA-CI; Guise et al. 2017b). These two checklists were developed by experts in systematic review methods to encourage research teams that conduct systematic reviews to engage in transparent, accurate, and comprehensive reporting of their review protocols.

Exhibit A.1. PRISMA-P elements

Element		Explanation	Section addressing
1a	Title	The title of this report is HomVEE Draft Handbook of Procedures and Evidence Standards. The title implies it is a systematic review handbook. The first sentence of the main text also states its relevance to a systematic review.	Front matter
1b	Update	This protocol updates standards and procedures guidance work previously published by HomVEE.	Chapter I, Section A
2	Registry	This review was not prospectively registered.	Not applicable
3a	Contact	Contact information appears on the title page, and stakeholders may contact the team through the website: https://homvee.acf.hhs.gov/.	Front matter
3b	Contributions	The ordering of the authors provides information on the relative contributions of each. Sama-Miller, as project director, is the guarantor of this work.	Front matter
4	Amendments	Version 2 amends original procedures and standards, including revisions specified in a pair of *Federal Register* notices: https://www.federalregister.gov/documents/2020/08/05/2020-17001/revised-procedures-and-standards-home-visiting-evidence-of-effectiveness-homvee-review https://www.federalregister.gov/documents/2020/08/05/2020-16992/updated-defintions-rules-and-procedures-related-to-model-versions-home-visiting-evidence-of (These *Federal Register* notices are also available on the HomVEE website.)	Chapter I, Section A
5a	Sources	This work was funded by the Office of Planning, Research, and Evaluation (OPRE), within the Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS).	Front matter, Chapter I
5b	Sponsor	This work was funded by OPRE, within ACF, HHS in partnership with HRSA.	Front matter, Chapter I
5c	Role of sponsor or funder	Staff from ACF and the Health Resources & Services Administration within HHS collaborated to develop content for and provide feedback on this protocol, and OPRE approved the draft. In 2009, an interagency work group of HHS staff helped shape the scope of the review. ACF, with input from HRSA, decides which home visiting models are prioritized for review each year.	Chapters I through III and Appendix B
6	Rationale	To help policymakers, program administrators, model developers, researchers, and the public identify well-	Chapter I

Element		Explanation	Section addressing
		designed research and understand which early childhood home visiting models are effective. One critical use of HomVEE results is to identify evidence-based models, a key requirement of eligibility for implementation with Maternal, Infant, and Early Childhood Home Visiting (MIECHV) Program funding.[61]	
7	Objective	This review seeks to summarize whether early childhood home visiting models for pregnant women and families with children from birth through age 5 are effective at improving outcomes in eight domains (see element 13).	Chapter I
8	Eligibility criteria	Research is eligible unless it is screened out for one of the following reasons: • The manuscript examines a home visiting model in a mandatory setting. • Home visiting was not the primary service delivery strategy studied in the intervention. (Models that provide services primarily in centers, with supplemental home visits, are excluded.) • The study that the manuscript examines did not use an eligible design (randomized controlled trials, quasi-experimental designs [single case, regression discontinuity and non-experimental comparison group],; see Chapter III). • The manuscript did not report results for an eligible target population: pregnant women or families with children whose ages range from birth to kindergarten entry (that is, up through age 5), and who are served in a developed-world context.[62] • The manuscript did not examine any findings in HomVEE's eight eligible outcome domains (listed in Exhibit I.1). • The manuscript did not examine a home intervention. • The manuscript was not published in English. • The study was published more than 20 years ago, unless it was submitted to the call for research or has already been reviewed by HomVEE.[63] • The manuscript did not present findings from primary research.	Chapter II

[61] For the purposes of the HomVEE review, this handbook uses the term "evidence-based model" to refer specifically to a model that meets HHS criteria developed based on statutory requirements in the authorizing legislation for the MIECHV Program. HomVEE recognizes that other systematic reviews may use different criteria to evaluate evidence of effectiveness. Thus, an evidence-based model in the context of HomVEE might or might not meet requirements for evidence of effectiveness according to other systematic reviews.

[62] HomVEE applies the term "developed-world context" to studies in countries that had high incomes in the year the manuscript was published, according to the World Bank Indicators list (World Bank 2020). For unpublished manuscripts, HomVEE will use the year the manuscript was submitted to the call for research.

[63] For models prioritized in 2018 and earlier, HomVEE also did a focused search reaching back to 1979. Because so few manuscripts published before 1979 related to models prioritized in recent years, starting with the 2019 review HomVEE limited the focused search to manuscripts reaching back to 1989 or later.

Element		Explanation	Section addressing
		For HomVEE, research evaluating the impact of a model feature or features is generally ineligible for review (see Exhibit III.1 for examples).	
9	Information sources	The review draws on database searches and a call for research.	Chapter II
10	Search strategy	The review used a modified version of the Peer Review of Electronic Search Strategies (PRESS) method (McGowan et al. 2016) to refine the database search terms in Exhibit II.2.	Chapter II, Section A
11a	Data management	The review uses a pair of databases (RefWorks and SharePoint) to catalog manuscripts and their corresponding studies as a management tool to track the literature search, screening, and review process.	Chapter II, Section A
11b	Selection process	HomVEE uses a multi-stage screening and prioritization process, and two reviewers examine each study.	Chapter II, Section A
11c	Data collection process	Data are recorded using a template based on that previously used by the What Works Clearinghouse, and by the HomVEE team under its initial standards, with updates to capture details needed for the new standards defined in this Handbook. HomVEE staff will conduct author queries to gather information not reported in the study.	Chapter II, Section B
12	Data items	Team members collect data at the manuscript, finding, and model levels.	Chapter II, Sections B and C
13	Outcomes and prioritization	HomVEE examines findings for outcomes in eight 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	Chapter III, Section A
14	Risk of bias in individual studies	Studies and findings are assigned an effectiveness rating based on several criteria, according to study design. Findings without sufficient causal validity will not be reported.	Chapter III, Sections B and C
15	Synthesis	Studies are grouped by home visiting model, and findings are summarized by model, qualitatively, by applying HHS criteria that designate which models are evidence based according to the quantity and type of favorable (statistically significant) findings.	Chapter II, Sections B and C
16	Meta-bias	HomVEE does not conduct meta-analyses nor assess meta-bases.	Not applicable
17	Confidence in cumulative evidence	The confidence in the evidence on each model will be summarized according to criteria defined by HHS for an evidence-based home visiting model.	Chapter II, Section B

Note: This exhibit follows Moher et al. (2015).

Exhibit A.2. PRISMA-CI methods elements not discussed in PRISMA-P

Element		Explanation	Section addressing
11a	Pathway complexity	This element (the presence of "complicated/multiple causal pathways, feedback loops, synergies, mediators, and moderators of effect" – Guise et al. 2017, p. 53) will vary across models. Therefore, HomVEE does not present an overall "analytic framework, causal pathway, or other graphical representation of the chain of evidence to illustrate the complexity of the causal pathway" (Ibid).	Not applicable
11b	Intervention complexity	This element varies across models and will be elaborated in an implementation report for each reviewed model. In these reports, HomVEE will detail available information on intervention components; the expected and actual frequency, duration, and intensity of service receipt; and the staff involved in service receipt.	Chapter II, Section C
11c	Population complexity	Manuscripts examining pregnant women and families with children from birth through age 5 are eligible for review if they meet other screening criteria. Each manuscript review further documents population characteristics.	Chapter I, Chapter II Section C
11d	Implementation complexity	This element varies vary across models and will be elaborated in an implementation report for each reviewed model. In these reports, HomVEE will detail available information on key implementation drivers.	Chapter II, Section C
11e	Contextual complexity	This element varies across models and studies and will be elaborated on in an implementation report for each manuscript rated moderate or high, and for each manuscript focused on an implementation study. In these reports, HomVEE will detail available information on the location of service receipt and local context.	Chapter II, Section C
11f	Timing	Services can occur for any length of time; however, the review focuses attention for newly reviewed research on manuscripts published in the past 20 years (applying a rolling window to each annual review cycle).	Chapter II, Section A
13	Summary measures	HomVEE reports effect sizes for each finding and average effect sizes by outcome domain and intervention.	Chapter II, Section C
14	Synthesis of results	Manuscripts are grouped into models and findings are summarized by model. Then, HomVEE reports all findings for a model and its related versions side by side.	Chapter II, Section C

Note: This exhibit follows Guise et al. (2017).

Appendix B

Outcomes Eligible for Review, and Assessable at Baseline

This page has been left blank for double-sided copying.

This appendix is organized by HomVEE outcome domain. For each domain, the appendix first lists measurement considerations. Then, it lists categories of outcomes, and sometimes specific measures of those outcomes, that HomVEE will categorize into each domain. When an assessment has scale or subscale scores that relate to several HomVEE domains, HomVEE usually places all outcomes and scales related to the assessment into a single domain. This eliminates the risk of individual subscale scores influencing the overall score, which would create unintended consequences for applying HHS criteria. An exception to this rule is assessments that have multiple scales but no overall score (such as the Protective Factors Survey). In that case, HomVEE sorts each scale into the domain to which it belongs. For each category or measure, it also specifies whether (or, under what circumstances) HomVEE considers the outcome to be assessable at baseline for the analyzed sample of families.

To develop these guidelines, HomVEE developed working decision rules for each domain and sorted categories of outcomes HomVEE had already seen to date. The contractor then took additional steps to confirm the guidelines. The process included examining documentation and validation studies from measure developers for applicable outcomes, identifying when and how some measures had been used in the past, and consulting with subject matter experts (SMEs) employed by the contractor that included a developmental psychologist and a pediatrician. Reviewers consult project leadership, SMEs, and HHS staff for guidance on outcomes not listed here, especially on measures that are very specific or highly technical.

A. Child development and school readiness

Outcomes in this domain include the child's social behaviors, attachment to a parent or caregiver, social-emotional or psychological development, and cognitive and academic development. Outcome measures in this domain include direct child assessments, reviews of school records, direct observations of children's behavior, and parent and teacher reports on standardized measures. Other outcome measures include parent and teacher reports on measures that are not standardized.

Note: If a parent or child reported that the child ran away from home, that measure would be reported in this domain. However, if the information were drawn from child welfare records, that measure would be listed under the reductions in child maltreatment domain.

1. Measurement considerations

Child mental and behavioral health belong in this domain. HomVEE categorizes measures of children's mental and behavioral health in the child development and school readiness domain, in contrast to the measures of physical health that are reported in the child health domain.

Formal, center-based child care. HomVEE categorizes measures of attendance at formal, center-based care in early childhood in this domain, and categorizes the direction of statistically significant impacts on such measures as having ambiguous direction.

Categorizing runaway information. If a parent or child reports that the child ran away from home, that measure would be reported in this domain. In contrast, if a child running away is information drawn from child welfare records, that measure would be listed under the reductions in child maltreatment domain.

Categorizing attachment measures. Attachment between parent and child is a dyadic concept that does not map precisely to one single outcome domain HomVEE focuses on, as specified in MIECHV

authorizing statute.[64] Therefore, if a measure of attachment examines child behavior, HomVEE places it in the child development and school readiness domain. Examples include attachment to the caregiver during infancy, engagement in a difficult task during toddler years, problem behaviors, and inhibitory control. In contrast, HomVEE places attachment measures that examine caregiver behavior (such as sensitivity and nurturance), as well as measures that are truly dyadic (such as the Dyadic Coercive Interactions measure in the Relationship Affect Coding System), in the positive parenting practices domain.

2. Guidelines on baseline assessability for eligible measures

Outcomes in the child development and school readiness domain are not assessable at baseline if any participants in the sample or subgroup being analyzed enroll prenatally. Given the age range (birth through five years) of children who are the focus of research HomVEE reviews, some outcomes in this domain will never be assessable at baseline (Table B.3). Other outcomes may be assessable at baseline if the entire analysis sample for a finding enrolls after the focal child's birth, if that sample is also entirely, at baseline, between the youngest and oldest ages within which outcome measure is assessed (Table B.4).

Table B.1. Child development and school readiness outcome measures considered unassessable at baseline by HomVEE

Outcome measure	Rationale for deeming measure unassessable at baseline
Academic attendance, performance, individualized instruction measures, and discipline	Not applicable for children before kindergarten entry. Includes: • School attendance or absence; • Focal child's attainment of high school diploma or GED; academic self-image measure, delayed entry into school; grade retention or placement; grade point average or course grade; Metropolitan Achievement Test; Metropolitan Readiness Test; Peabody Individual Achievement Test; Stanford Early Achievement Test; Test of Early Reading Ability; • Receipt of special education, remedial, or therapeutic services • Sent to principal's office
Achenbach Youth Self Report of Problem Behaviors, including internalizing and externalizing behaviors	Not measured for infants, toddlers, or preschool age children (but, other assessments of internalizing and externalizing behaviors are appropriate for younger children).
Antisocial Process Screening Device (APSD)	Measured for children ages 6 to 13 years
Child Behavior Rating Scale (CBRS)	Measured for children ages 6 through 18 years
Child ran away from home (parent or child reported)	Not generally asked or recorded for children ages 5 years or younger
Child's risky behavior as youth or juvenile (including sexual behavior, parenting as a teen, substance use)	Not generally asked or recorded for children ages 5 years or younger
Halstead-Reitan Neuropsychological Test Battery	Measured for children ages 5 to 16 years
Kerns Security Scale	Measured for children age 6 years and older

[64] Social Security Act, Section 511 [42 U.S.C. 711]

Table B.2. Baseline assessability of other outcome measures in HomVEE's child development and school readiness domain

Measurement concept	Youngest age to assess	Oldest age to assess
Psychosocial development		
Attachment		
Attachment Q-Set Scale (AQS)	1 year old	5 years old
Strange Situation Procedure	9 months	18 months
Socio-emotional/psychological development, behavior, and mental health		
Achenbach Child Behavior Checklist (CBCL), including internalizing and externalizing behaviors	18 months	18 years old
Adaptive Social Behavior Inventory (ASBI)	3 years old	5 years old
Behavior Assessment System for Children (BASC)-2, including internalizing and externalizing behaviors	2 years old	25 years old
Child Behavior Questionnaire (CBQ) inhibitory control	3 years old	7 years old
Child's crying and irritability	3 days	6 months
Emotion Regulation Checklist	3 years old	11 years old
Eyeberg Child Behavior Inventory (ECBI)	2 years old	16 years old
Fussiness Rating Scale	Birth	1 year old
Hightower Teacher-Child Rating Scale (HTC)	Pre-K	3rd grade
Infant-Toddler Social and Emotional Assessment (ITSEA), or Brief ITSEA (BITSEA)	1 year old	3 years old
Infant Temperament Questionnaire (ITQ)	Birth	36 months
Mastery motivation	15 months	30 months
Physiologic measures of regulation and stress response (such as skin conductance level and respiratory sinus arrhythmia)	Request SME guidance about age for the specific measure	Request SME guidance about age for the specific measure
Social Skills Rating System (SSRS)	3 years old	18 years old
Cognitive development		
Language development		
Bracken Basic Concept Scale (BBCS-R)	3 years old	6 years old and 11 months
Expressive One-Word Picture Vocabulary Test	2 years old	80+ years old
Fluharty-2 Preschool Speech and Language	3 years old	6 years old and 11 months
Language Acquisition Quotient-Zimmerman Preschool Language Scale	Birth	6 years old and 11 months
MacArthur Communicative Development Inventory (CDI)	8 months	37 months
MacArthur Story Stem Battery (MSSB)	3 years old	7 years old
Peabody Picture Vocabulary Test (PPVT)	2 years old and 6 months	90 years old
Woodcock-Johnson Tests of Achievement (WJ-III, WJ-IV Oral Language, WJ-Cognitive Oral Vocabulary and Picture Recognition)	2 years old	80 years old
Other cognitive development (some assessments include physical development items)		
Ages & Stages Questionnaire (ASQ)	1 month	5.5 years
Bayley Scales of Infant Development (BSID)	1 month	42 months
Cooperative Preschool Inventory (CPI) also includes social behavior components	3 years old	6 years old
Denver Developmental Screening Tests	Birth	6 years old
Developing Skills Checklist	4 years old	End of kindergarten

Measurement concept	Youngest age to assess	Oldest age to assess
Developmental Profile II (DPII)	Birth	12 years old and 11 months
Kaufman Assessment Battery for Children (ABC)	3 years old	18 years old
Leiter International Performance Scale–Revised	3 years old	75+ years old
Stanford-Binet Intelligence Test	2 years old	85+ years old
Wechsler Preschool and Primary Scale of Intelligence (WPPSI)	2 years old and 6 months	7 years old and 7 months

SME= subject matter expert.

B. Child health

Measures of a child's growth, physical health, and use of health services (such as immunizations) are all included in this domain. Outcome measures in this domain are birth outcomes and counts of health care service use, which are extracted from medical records. Other outcome measures in this domain are based on parent reports about children's health and use of health care services. It is also important to note that some unfavorable or ambiguous outcomes, such as number of days hospitalized, may be due to increased access to health care. In other words, families' participation in the home visiting program may have increased the likelihood that they would receive needed health care services.

1. Measurement considerations

Diet and feeding measures belong in this domain. One related outcome, the Breastfeeding Self-Efficacy Scale, belongs in the positive parenting practices domain instead because it measures the parent's attitudes.

Child mental and behavioral health do not belong in this domain. HomVEE categorizes measures of children's mental and behavioral health in the child development and school readiness domain, in contrast to the measures of physical health that are reported in the child health domain.

Health care encounters due to injuries and ingestions do not belong in this domain. Most health care encounters for children belong in the child health domain. However, health care encounters that may occur specifically as a result of child maltreatment, such as treatment for injuries or ingestions, are placed in the reductions in child maltreatment domain.

Health insurance coverage does not belong in this domain. HomVEE places access to health insurance, for both the child and mother, in the family economic self-sufficiency domain.

2. Guidelines on baseline assessability for eligible outcomes

Outcomes in the child health domain generally are not assessable at baseline if any participants in the sample or subgroup being analyzed enroll prenatally. The exception is measures of prenatal health of the focal child. Given the age range (birth through age 5 years) of children who are the focus of research HomVEE reviews, some outcomes in this domain will never be assessable at baseline (Table B.1). Other outcomes may be assessable at baseline if the entire analysis sample for a finding enrolls after the focal child's birth, and if that sample is also entirely, at baseline, between the youngest and oldest ages within which the outcome measure is assessed (Table B.2). The highly specific nature of some child health measures means that HomVEE will consult an SME, such as a pediatrician or other physician for additional guidance as needed.

Table B.3. Child health outcome measures considered unassessable at baseline by HomVEE

Outcome measure	Rationale for deeming measure unassessable at baseline
Birth outcome measures (weight, length, Apgar score, gestational age)	These are one-time measures that could not be assessed at both baseline and follow-up.
Infant or child mortality	Families in this situation would leave the home visiting evaluation.

Table B.4. Baseline assessability of other outcome measures in HomVEE's child health domain

Measurement concept	Youngest age to assess	Oldest age to assess
Growth and weight		
Child's weight for age, as percentile or Z score	2 months	18 years
Child's length for height, as percentile or Z score	2 months	18 years
Child's BMI, as percentile or Z score	2 years	18 years
Physical health		
Child's physical illness (acute)	Birth	18 years
Child's prenatal health problem	Prenatal	Prenatal
Child health problems (chronic)	As young as birth, but depends on the problem being measured; request SME guidance	18 years
Specific health indicators (cotinine levels, cortisol levels, telomere length)	Request SME guidance, including on period of potential primary exposure (including prenatal and postnatal) versus secondary	Request SME guidance
Health services usage		
Child's immunizations	Request SME guidance because age-dependent, but generally age 2 years	18 years
Well child check-ups	Within a few days of birth	18 years
Use of health services, including general hospital/emergency services for child (not specific to injury/ingestion)	As young as birth, but depends on the problem being measured; request SME guidance	18 years
Other child health measures		
Parent's choices in feeding child: Breastfeeding, formula, or water	Birth	12 months
Pediatric Quality of Life Inventory	2 years	18 years
Parent's choices in feeding child: Juices or solid foods	Generally, 3 to 6 months, but request SME guidance	18 years

Note: SME= Subject Matter Expert

C. Family economic self-sufficiency

Outcomes in this domain measure a family's economic well-being, including income and earnings, receipt of means-tested public assistance, and access to resources such as housing and transportation. The family economic self-sufficiency outcomes also measure employment and educational enrollment or attainment, as well as other sources of support, such as child support from a noncustodial parent. Measures of the mother's partnership status (married, cohabiting, and so on) are ineligible for review. Outcome measures in this domain include measures of public assistance receipt that are based on government administrative records and maternal self-reports of service receipt and economic outcomes.

1. Measurement considerations

Primary caregiver and overall household economic well-being outcomes belong in this domain. These include primary caregivers' educational attainment and enrollment, and their income and earnings. Eligible outcomes also include overall household income, access to transportation, other sources of financial support such as child support from a noncustodial parent, and receipt of means-tested public assistance.[65] By extension, including measures of the father's or mother's current partner's socioeconomic status (such as that person's education, employment, or earnings), are ineligible for review *unless* (1) the manuscript reports these same outcomes for the mother as well and (2) the father or partner is coresident (so that HomVEE can assess the overall situation of the household).

In-kind support that the primary caregiver received is categorized in this domain, whereas social/emotional support to the mother belongs in the maternal health domain. HomVEE generally characterizes more support as favorable.[66]

Measures of the mother's partnership status (married, cohabiting, and so forth) are not eligible for review, because they are not clear indicators of family economic self-sufficiency.

2. Guidelines on baseline assessability for eligible outcomes

Eligible outcomes in this domain (listed below) are assessable at baseline unless they can only be assessed after the focal child has been born (denoted in the list below *in italics with an asterisk**).

- Economic well-being measures
  - Household income
  - Earnings of primary caregiver
  - Poverty level according to federal thresholds
  - Other socioeconomic measures
    - International socioeconomic measures (for example, Elley-Irving Socio-Economic Index)
    - Neighborhood Disadvantage Index

[65] The favorability of positive or negative findings in this area can depend on context or other factors. In some cases, the HomVEE team will confer with subject matter experts to determine whether a finding is favorable, or unfavorable or ambiguous.

[66] The favorability of positive or negative findings in this area can depend on context or other factors. In some cases, the HomVEE team will confer with subject matter experts to determine whether a finding is favorable, or unfavorable or ambiguous.

    - Hollingshead Four Factor Index of Socioeconomic Status
    - Food insecurity
  - Employment status, duration for primary caregiver
  - Income, earnings, or education of child's father or mother's current partner **only if** the manuscript also reports findings about the same outcomes for the mother and the father or partner is coresident (that is, the study results approximate the overall situation of the household).
    - *Although these outcomes are assessable at baseline, fathers also must reside with the child in order to use father's SES measures to establish baseline equivalence when HomVEE reviews the study, and other criteria about establishing baseline equivalence also apply, as described in Exhibit III.11 of the handbook.*
- Education or training enrollment or attainment for primary caregiver
- Means-tested assistance measures for household or focal child[67]
  - Temporary Assistance for Needy Families (TANF)
  - Supplemental Nutrition Assistance Program (SNAP)
  - **Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)*
- Health insurance measures
  - Medicaid
  - **Children's Health Insurance Program*
  - **Child's health insurance status*
  - Mother's health insurance status
- **Child support (from noncustodial parent)*
- In-kind support from family, such as helping mother to provide child care
- Family resources
  - Housing or homelessness
  - Transportation access
  - Family Resources Scale

D. Linkages and referrals

These measures assess whether the home visiting model has referred a family to services such as early intervention, child care, or public benefit programs. Outcome measures in this domain include reviews of home visitor, medical, or school records for indications that the child or family had received a referral to other services in the community, as well as parent reports of receiving a referral and being aware of other services in the community.

For this domain only, HomVEE includes outcomes measured at the provider and family levels. For example, HomVEE would include the number of referrals that a home visitor or other service provider

[67] The favorability of positive or negative findings in this area can depend on context or other factors. In some cases, the HomVEE team will confer with subject matter experts to determine whether a finding is favorable, or unfavorable or ambiguous.

gave to families. This is consistent with the benchmark areas in the MIECHV authorizing statute, which include coordination and referrals for other community resources and supports.

1. Guidelines on baseline assessability for eligible outcomes

Outcomes in the linkages and referrals domain are not assessable at baseline because referrals, for many home visiting interventions, are a direct service of the intervention and are not logical to measure before services have begun. Eligible outcomes include the following:

- Referral to parent's education-related, vocational or employment-related services
- Referral to public benefit programs
- Referral to medical services
- Referral to mental health services
- Referral for early intervention or to services for child's disability
- Referral for child's education-related services
- Referral for child care services
- Referral to services for immigrants

Outcomes in this domain also include measures of linkages of study participants to resources in their communities (for example, what resources participants knew about or accessed), even if those linkages are indirect rather than the specific result of a referral by a home visitor or other service provider.

E. Maternal health

Maternal health involves the mother's health status (during or after pregnancy), including mental and behavioral health, stress levels, health-related habits such as nutrition and sexual health, and measures of social support and other protective factors. Outcome measures in this domain include health care service receipt outcomes, which are extracted from medical records, as well as standardized and unstandardized parent self-report measures.

1. Measurement considerations

Maternal health involves the mother's health status (during or after pregnancy), including mental and behavioral health, stress, and health-related habits such as nutrition and sexual health. Receipt of health services is in this domain; the mother's health insurance status is in the family economic self-sufficiency domain.

Social/emotional support to the mother belongs in this domain, whereas in-kind support that the primary caregiver received is in the family economic self- sufficiency domain. HomVEE generally characterizes more support as favorable.[68]

[68] The favorability of positive or negative findings in this area can depend on context or other factors. In some cases, the HomVEE team will confer with subject matter experts to determine whether a finding is favorable, or unfavorable or ambiguous.

2. Guidelines on baseline assessability for eligible outcomes

Outcomes in the maternal health domain are always assessable if they measure a general aspect of women's health (such as a mother's substance use) that is not contingent upon pregnancy, parenthood or the birth/presence of a child. Outcomes that are contingent upon the birth of a child are not assessable at baseline if families are enrolled in the study before the birth of the focal child. Table B.5 provides details on baseline assessability in each case.

Table B.5. Baseline assessability of outcome measures in HomVEE's maternal health domain

Outcome measure	Description and notes	Assessable at baseline when family enrolled…	
		…during the mother's pregnancy	…at or after focal child's birth
Mother's physical health			
Maternal receipt of general health services	Measures may include number and frequency of visits with a provider for general physical and behavioral health services. Note: Measures of mother's health insurance status instead belong in the family economic self-sufficiency domain (see next section for discussion of measures of prenatal care).	Yes	Yes
Health status during pregnancy	Includes maternal receipt of prenatal services; mother's gestational health status, and specific diagnoses measured in pregnancy (such as gestational diabetes); and the Pregnancy Risk Assessment Monitoring System (PRAMS)	Yes	No, except perhaps PRAMS (which can be assessed shortly after birth)
Birth outcomes	One-time measures of the mother's health at birth, including maternal mortality	Not applicable	No
Pregnancies, births, miscarriages, or abortions after the birth of the focal child	The favorability of positive or negative findings in this area can depend on context or other factors. In some cases, the HomVEE team will confer with subject matter experts to determine whether a finding is favorable, or unfavorable or ambiguous.	No	No
Mother's mental health, behavioral health, and habits			
Depression and anxiety	Includes Center for Epidemiologic Studies - Depression (CES-D) assessment	Yes, unless measure is specific to postpartum mood disorders	Yes
Diet and nutrition		Yes	Yes
Maternal mastery/self-esteem/empowerment/self-efficacy/resiliency	Includes Family Crisis Oriented Personal Evaluation Scales (F-COPES)	Yes	Yes

Outcome measure	Description and notes	Assessable at baseline when family enrolled…	
		…during the mother's pregnancy	…at or after focal child's birth
Mental health assessments	Includes Composite International Diagnostic Interview (CIDI) assessment for mental disorders and Mental Health Inventory (MHI), and Structured Clinical Interview for Diagnostic and Statistical Manual (SCID), and assessments of internalizing and externalizing behaviors	Yes	Yes
Problem Oriented Screening Instrument for Teenagers (POSIT)	Assumes the teenagers being assessed are mothers in the home visiting program (if they are children formerly served by a home visiting model, this is a child health measure)	Yes	Yes
Sexual health		Yes	Yes
Substance use	Includes alcohol, cigarettes, and drugs	Yes	Yes
Maternal stress	Includes Parenting Stress Index (PSI) and Perceived Stress Scale (PSS)	Yes, for measures not restricted to parents	Yes, PSI is assessable as soon as child is are 1 month old
Maternal social support, coping skills, and protective factors	Includes Community Life Skills Scale (CLSS), Inventory of Socially Supportive Behaviors (ISSB) Maternal Social Support Index (MSSI), Protective Factors Survey (PFS), Social Provision Scale	Yes	Yes

F. Positive parenting practices

Outcomes in this domain include knowledge of child development, safety practices, supportive behavior and engagement with the child, promotion of learning and child development, disciplinary practices, and general parenting practices such as bedtime routines. Outcome measures in this domain include observational measures of parent-child interactions or the home environment. For some measures, parent-child interactions are videotaped and then coded at a later time. For others, live coding is completed during an observation of the parent and child in the home environment. Many studies also use outcome measures based on parent self-reports of parenting attitudes and practices.

1. Measurement considerations

Categorizing attachment measures. Attachment between parent and child is a dyadic concept that does not map precisely to one single outcome domain HomVEE focuses on, as specified in statute. Therefore, HomVEE places attachment measures that examine caregiver behavior (such as sensitivity and nurturance), as well as measures that are truly dyadic (such as the Dyadic Coercive Interactions measure in the Relationship Affect Coding System), in the positive parenting practices domain. In contrast, if a measure of attachment examines child behavior, HomVEE places it in the child development and school

readiness domain. Examples include attachment to the caregiver during infancy, engagement in a difficult task during toddler years, problem behaviors, and inhibitory control.

2. Guidelines on baseline assessability for eligible outcomes

Outcomes in the positive parenting practices domain generally are not assessable at baseline if the measure presumes the presence of a child, such as in a measure of parent-child interaction. If the family enrolls after the birth of the focal child, most parenting outcomes are assessable at baseline if that sample is also entirely, at baseline, within the youngest and oldest ages assessable by that measure (Table B.6, last column). The highly specific nature of some parenting measures means that HomVEE will consult an SME, such as a psychologist or child development expert, for additional guidance as needed.

Outcomes that are measured only once (such as the D.O.T.S Emotion Coding System, which was developed to be administered when children are age 24 months and was not tested with other ages) are not assessable at baseline because it would not be possible to assess the outcome at both baseline and follow-up.

Table B.6. Baseline assessability of outcome measures in HomVEE's positive parenting practices domain

Outcome	Assessable during pregnancy, about focal child	Assessable only after focal child's birth (for specific baseline child age)
Knowledge of child development		
Knowledge of Infant Development Inventory (KIDI)	Yes, during third trimester	Yes (until age 2 years)
Parent Development Interview-Revised (PDI)	No	Yes (beginning in infancy)
Toddler Care Questionnaire (TCQ)	No	Yes (12 through 36 months)
Parenting safety and home environment		
General home environment and safety	Yes	Yes
Home Observation for Measurement of the Environment (HOME)	No	Yes (beginning at birth)
Safe sleep practices	Knowledge about practices: Yes, during third trimester *Implementation of practices:* No	Yes (birth through 12 months)
Family Environment Scale (FES)	Yes	Parent or family member older than 11 is respondent
Parent's engagement with child and supportive behavior		
Atypical Maternal Behavior Instrument for Assessment and Classification (AMBIANCE)	No	Yes (ages 12 through 24 months; adapted AMBIANCE can be measured as young as 4 months)
Nursing Child Assessment Teaching Scale	No	Yes (birth to 3 years)
Keys to Interactive Parenting Scale (KIPS)	No	Yes (beginning at 2 months)
Maternal engagement/relationship with child	No	Yes (request SME guidance about age given specific measure)
Family Assessment Device (McMaster), or McMaster Clinical Rating scale of family functioning (observation), or McMaster structured interview of family functioning	No	Yes (beginning at birth; respondent [caregiver] should be 12 years or older)
Father's contact with child	No	Yes (beginning at birth)

Outcome	Assessable during pregnancy, about focal child	Assessable only after focal child's birth (for specific baseline child age)
Parent-child interaction - parent behavior/ responsiveness	No	Yes (request SME guidance about age given specific measure)
Parent-Child Activities Scale (PCAS)	No	Yes (beginning at 6 months)
Relationship Affect Coding System (RACS)	No	Yes (beginning at 2 years)
Relationship Process Code	No	Yes (beginning at 2 years)
Verbal encouragement	No	Yes (request SME guidance about age given specific measure)
Perceptions of parenting role		
Breastfeeding Self-Efficacy Scale (BSES)	No	Yes (birth through 12 months) Other measures of breastfeeding are in the child health domain
Parental Locus of Control (PLOC)	No	Yes (beginning at birth)
Parenting Sense of Competence (PSOC)	No	Yes (beginning at birth)
General parenting practices		
Adult-Adolescent Parenting Inventory (AAPI)	Yes (pre-parent should be at least 12 years old)	Yes (parent should be at least 12 years old)
Child exposure to television or books	No	Yes (usually, for children beginning at age 12 months; consult SME for samples enrolled younger than 12 months)
Child Rearing Practices Report (CRPR)	No	Yes (beginning at age 2 years)
Family Involvement Questionnaire	No	Yes (preschool through grade 1)
Healthy Families Parenting Inventory (HFPI)	No	Yes (request SME guidance about age)
Observational Record of the Caregiving Environment (ORCE)	No	Yes (6, 15, 24, and 36 months)
Parent-Infant Interaction Observation Scale	No	Yes (2 through 7 months)
Parental disciplinary actions towards child	No	Yes (request SME guidance about age given specific measure) Exception: all measures from the Conflict Tactics Scale-Parent Child fall under the reductions in child maltreatment domain
Parent Behavior Checklist	No	Yes (1 through 4 years)
Parenting Scale (PS)	No	Yes (18 months through age 5 years)
Planned Activities Training (PAT) checklist	No	Yes (8 months to 5 years)
Promotion of learning, language, and development	No	Yes (usually, for children beginning at 12 months; consult SME for samples enrolled younger than 12 months)
Routines and bedtime	No	Yes (usually, for children beginning at 12 months; consult SME for samples enrolled younger than 12 months)
Tummy time (whether administered by parent)	No	Yes (birth to 6 months)

G. Reductions in child maltreatment

Outcomes in this domain include measures and assessments related to child maltreatment. Outcome measures include evidence of substantiated child maltreatment from administrative records and counts taken from medical records of encounters with health care providers for injuries or ingestions. Encounters with health care providers may include physician visits, emergency room visits, or hospitalizations. Parents in home visiting programs may be encouraged to use health care services more often, such as for well-child care visits. In addition, families' patterns of health care use may change after enrollment in a home visiting program. For example, if a program connected families with primary care physicians, they might reduce their use of the emergency room for health care. Therefore, in the HomVEE review, only health care encounters that may occur as a result of child maltreatment, such as treatment for injuries or ingestions, are included in the child maltreatment domain.

There is some concern that counts of child maltreatment reports may not be accurate indications of the incidence of maltreatment. For example, participation in home visiting programs increases surveillance of families and may result in increased reports of child maltreatment. Therefore, this review includes only substantiated reports of child maltreatment as an outcome measure; outcome measures based on unsubstantiated reports are excluded. HomVEE also includes child welfare outcomes such as placement outside the home.

HomVEE has classified the Conflicts Tactics Scale-Parent Child (CTS-PC), a measure that assesses neglectful, psychologically aggressive, and abusive parenting behavior, as an assessment that measures child maltreatment.

1. Measurement considerations

HomVEE includes only substantiated reports of child maltreatment and child welfare measures such as custody loss and placement outside the home; outcome measures based on unsubstantiated reports are ineligible for review. There is some concern that counts of child maltreatment reports may not be accurate indications of the incidence of maltreatment. For example, participation in home visiting programs increases surveillance of families and may result in increased reports of child maltreatment.

Only health care encounters that may occur specifically as a result of child maltreatment, such as treatment for injuries or ingestions, are included in the reductions in child maltreatment domain. Encounters with health care providers may include physician visits, emergency room visits, or hospitalizations. Parents in home visiting programs may be encouraged to use health care services more often, such as for well child care visits. In addition, families' patterns of health care use may change after enrollment in a home visiting program. For example, if a program connects families with primary care physicians, families may reduce their use of the emergency room for health care. Therefore, HomVEE places other health care encounter measures in the child health domain.

Categorizing runaway information. If a child running away is measured in child welfare records, that measure would be listed here. In contrast, if a parent or child reports that the child ran away from home, that measure would be reported in the child development and school readiness domain.

2. Guidelines on baseline assessability for eligible measures

Table B.7. Baseline assessability of outcome measures in HomVEE's reductions in child maltreatment domain

Outcome measure	Description and notes	Assessable at baseline when family enrolled…	
		…during prenatal period for focal child	…at or after focal child's birth
Substantiated child abuse or neglect cases	Measures may include overall, by type of abuse or neglect, or by timing of abuse or neglect. Also includes measures of family reunification	No	Yes, consult SME for guidance given child's age at enrollment.
Permanency	Refers to the permanency and stability of a child's living situation (in-home or in foster care) and includes the continuity and preservation of family relationships and connections. This includes measures of custody loss due to abuse or neglect, and measures of placement outside the home	No	Yes, consult SME for guidance given child's age at enrollment.
Unsubstantiated child abuse or neglect cases	Ineligible for review	Not applicable	Not applicable
Health care encounter due to injury or ingestion	May be described as emergency room visit or hospital visit; measures may include overall incidence, timing, and frequency within a follow-up period	No	Yes. Reviewers will assume equivalence if enrollment occurred at birth, during mother's postpartum hospital stay.
Health care encounter for other reasons	Please see child health outcome domain	Not applicable	Not applicable
Conflict Tactics Scale–Parent Child (CTS-PC)	HomVEE includes all items from this assessment in the reductions in child maltreatment domain in order to make domain classification consistent for summary measures from the assessment. Note: This is a secondary measure (not normed)	No	Yes Reviewers will assume equivalence if enrollment occurred at birth during mother's postpartum hospital stay.
Child Abuse Potential Inventory (CAPI)	Note: This is a secondary measure (not normed)	No	Yes Reviewers will assume equivalence if enrollment occurred at birth during mother's postpartum hospital stay.
Child ran away from home (CPS reported)		No	No

H. Reductions in juvenile delinquency, family violence, and crime

In this domain, outcomes may include domestic and family violence, interaction with the justice system by the mother or by a youth who received home visiting services during early childhood, or school suspensions or expulsions for one of these youth. Outcome measures in this domain include the incidence of parent and youth antisocial behavior, based on archived data from state records, as well as parent, teacher, and youth self-report of antisocial behaviors. For example, HomVEE places the Conflict Tactics Scale (CTS), a measure that assesses family violence, intimate partner violence and child maltreatment, in this domain.

1. Guidelines on baseline assessability for eligible measures

Table B.8. Baseline assessability of outcome measures in HomVEE's reductions in juvenile delinquency, family violence, and crime domain

Outcome measure	Description and notes	Assessable at baseline when…	
		…family enrolled during prenatal period for focal child	…family enrolled at or after focal child's birth
Parent's intimate partner violence (IPV), family violence, or domestic violence	Measures may include physical or psychological violence, with parent as victim or perpetrator; may also include restraining order	Yes	Yes
Parental interaction with justice system	Measures may include arrests, convictions, incarcerations, and measures of specific offenses.	Yes	Yes
Focal child's interaction with justice system as a juvenile	Measures may include arrests, convictions, and measures of specific offenses, and whether the youth was ever a "person in need of supervision"	No	No
Focal child's school suspension or expulsion	None	No	No
Focal child's risky behavior as a youth	Classified under child health domain	Not applicable to this domain	Not applicable to this domain

This page has been left blank for double sided copying.

Appendix C

Standards for Regression Discontinuity Designs

This page has been left blank for double sided copying.

This appendix replicates the What Works Clearinghouse Version 4.1 standards for research with this design, except for minor wording changes to tailor them to the HomVEE context.[69]

Researchers use regression discontinuity designs (RDDs) when interventions are made available to individuals or groups on the basis of how they compare with a cutoff value on some known measure. Sample members may be assigned, for example, to a program if they score below a cutoff value on a given assessment.[70] The variable used to assign participants to the intervention is commonly referred to as the "forcing," "assignment," or "running" variable.

The effects provide consistent estimates of the local average impacts and are comparable with traditional group design trials. Under typical RDD methodology, the effect of an intervention is estimated as the difference in mean outcomes between intervention and comparison group members at the cutoff, adjusting statistically for the relationship between the outcomes and the variable used to assign participants to the intervention. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and the difference in these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. Stated differently, an effect is said to have occurred if there is a "discontinuity" in the two regression lines at the cutoff. This estimate pertains to average intervention effects for participants right at the cutoff. RDDs generate asymptotically unbiased estimates of the effect of an intervention if the relationship between the outcome and forcing variable is modeled appropriately (defined in standard 4 next) and the forcing variable was not manipulated, either behaviorally or mechanically, to influence assignment to the intervention group.

This appendix presents criteria under which estimates of effects from RDD studies can be rated high and the conditions under which they can be rated moderate. These standards apply to both "sharp" and "fuzzy" RDDs, defined in Section C. We provide standards for studies that report a single RDD impact (Section C), standards for studies that report multiple impacts (Section D), and standards for studies that report pooled or aggregate impacts (Section E). As is the case in RCTs, clusters of students—such as schools, classrooms, or any other group of multiple individuals that have the same value of the assignment variable—might be assigned to intervention and comparison groups, and so we provide standards for cluster-assignment studies (Section F). While the standards are focused on assessing the causal validity of impact estimates, we also describe two reporting requirements (Sections G and H) focused on reporting accurate standard errors.

A. Assessing whether a manuscript about a study is eligible for review as a regression discontinuity design

A manuscript is eligible for review under RDD standards if it meets the following criteria:

- Treatment assignments are based on a numerical forcing variable; participants with numbers at or above a cutoff value, or at or below that value, are assigned to the intervention group, whereas participants with scores on the other side of the cutoff are assigned to the comparison group. For example, an evaluation of a home visiting program could be classified as an RDD if families with a Family Stress Checklist (FSC) score at or above 25 are admitted to the program and families with an FSC score below 25 are not. As another example, a study examining the impacts of a home visiting

[69] What Works Clearinghouse. (2020). Handbooks and Other Resources: Procedures and Standards Handbooks. Retrieved June 4, 2020, from https://ies.ed.gov/ncee/wwc/handbooks.

[70] Generally, groups of sample members may also be sorted on either side of a cut point in a regression discontinuity design, but assignment of groups is exceptionally rare in the home visiting literature.

program on improving families economic self-sufficiency could be considered an RDD if only families with a poverty index score at or below a threshold are admitted to the program and families with a poverty index score above a threshold are not. In some instances, RDDs may use multiple criteria to assign the treatment to participants. For example, a family may be assigned to a home visiting program if the family's FSC score is above 25 or the depression assessment score is above a threshold (for example, a summary score of depressive symptoms is above 8 on the PHQ-8 scale). Studies that use multiple assignment variables or cutoffs with the same sample are eligible for review under these standards only if they use a method described in the literature (for example, in Reardon and Robinson [2012] or Wong, Steiner, and Cook [2013]) to reduce those variables to a single assignment variable or analyze each assignment variable separately. If a study does not do this (for example, if it uses the response surface method described by Reardon and Robinson [2012]), then a manuscript about it is not currently eligible for review under these standards. As with randomized controlled trials (RCTs), noncompliance with treatment assignment is permitted, but the manuscript about the study must still meet the criteria outlined in this appendix to be eligible for a rating of high or moderate.

- The forcing variable is *ordinal*—that is, it has a unique ordering of the values from lowest to highest—and includes a minimum of four or more unique values below the cutoff and four or more unique values above the cutoff. This condition is required to model the relationship between the outcomes and the forcing variable. The forcing variable must never be based on nonordinal categorical variables, such as sex or race. The analyzed data must also include at least four unique values of the forcing variable below the cutoff and four unique values above the cutoff. This is required for eligibility because at least eight data points are required to credibly select bandwidths or functional forms for the relationship between the outcome and the forcing variable.
- The study must not have a *confounding factor* as defined for RCTs and non-experimental comparison group designs (NEDs) in Chapter III. As defined there, for HomVEE, a **confounding factor** is any observed factor that is not completely aligned with either the intervention or comparison group. In particular, the cutoff value of the forcing variable must not be used to assign members of the study sample to interventions other than the one being tested. For example, the income cutoff for determining whether a family qualifies for the Special Supplemental Program for Women, Infants, and Children (WIC) cannot be the basis of an RDD because WIC receipt could affect maternal and child health outcomes that are of interest to HomVEE. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions. A study can examine the combined impact of two or more interventions that all use the same cutoff value; in that case, the manuscript about the study can be eligible for review as an RDD, but the causal statements made must be about the combined impact because the causal effects of each individual intervention cannot be isolated.
- The forcing variable used to calculate impacts must be the *actual* forcing variable, not a proxy or estimated forcing variable. A variable is considered to be a proxy if its correlation with the actual forcing variable is less than 1.

If a study claims to be based on an RDD but does not have these properties, then any manuscripts about the study are not eligible for review as an RDD.

B. Possible ratings for studies using regression discontinuity designs

Once a study is determined to be an RDD, findings within a manuscript about the study can receive one of three ratings based on the set of criteria described below and summarized in Table C.1. The manuscript itself receives the highest rating of any finding within it.

- **High**. To qualify, the manuscript must completely satisfy each of the five individual standards listed in Table C.1.
- **Moderate**. To qualify, the manuscript must at least partially satisfy each of the following standards: 1, 4, 5, and either 2 or 3.
- **Low**. A manuscript about an RDD study will receive this rating if it does not at least partially satisfy any of standards 1, 4, or 5, or does not at least partially satisfy both standards 2 and 3.

Table C.1. Regression discontinuity design manuscript ratings

Standard	To be rated Meets WWC RDD Standards Without Reservations, a manuscript about an RDD study must:	To be rated Meets WWC RDD Standards With Reservations, a manuscript about an RDD study must:
1. Integrity of the forcing variable	Completely satisfy this standard.	Partially satisfy this standard.
2. Sample attrition	Completely satisfy this standard.	Partially satisfy at least one of these two standards.
3. Continuity of the relationship between the outcome and the forcing variable	Completely satisfy this standard.	Partially satisfy at least one of these two standards.
4. Functional form and bandwidth	Completely satisfy this standard.	Partially satisfy this standard.
5. Fuzzy RDD	Completely satisfy this standard.	Partially satisfy this standard.

C. Standards for a single regression discontinuity design impact

The standards presented in this section focus on assessing the causal validity of the impact of a single discontinuity in a single ordinal forcing variable on a single outcome. Section D describes how to apply these standards in studies with multiple outcomes or samples. Section E describes how to apply these standards in studies with pooled or aggregate impacts.

Standard 1: Integrity of the forcing variable

A key condition for an RDD to produce consistent estimates of effects of an intervention is that there was no systematic manipulation of the forcing variable. This situation is analogous to the nonrandom manipulation of intervention and comparison group assignments under an RCT. In an RDD, manipulation means that scores for some participants were systematically changed from their true obtained values to influence treatment assignments and the true obtained values are unknown. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to inconsistent impact estimates.

Manipulation is possible if "scorers" have knowledge of the cutoff value and have incentives and an ability to change unit-level scores to ensure that some participants are assigned to a specific research condition. Stated differently, manipulation could occur if the scoring and treatment assignment processes

are not independent. It is important to note that manipulation of the forcing variable is *different* from treatment status noncompliance, which occurs if some intervention group members do not receive intervention services or some comparison group members receive embargoed services.

The likelihood of manipulation will depend on the nature of the forcing variable, the intervention, and the study design. For example, manipulation is less likely to occur if the forcing variable is a standardized assessment than if it is a family assessment conducted by researchers or home visitors who also have input into treatment assignment decisions. Manipulation is also unlikely in cases where the researchers determined the cutoff value using an existing forcing variable, for example, a score from a test that was administered prior to the implementation of the study.

In all RDD studies, the *integrity of the forcing variable* should be established institutionally, statistically, and graphically.

- Criterion A. The institutional integrity of the forcing variable must be established by an adequate description of the scoring and treatment assignment process. This description must indicate the forcing variable used; the cutoff value selected; who selected the cutoff—for example, researchers and model developers; who determined values of the forcing variable—for example, who scored an assessment; and when the cutoff was selected relative to determining the values of the forcing variable. This description must show that manipulation was unlikely because scorers had little opportunity or little incentive to change "true" obtained scores in order to allow or deny specific participants access to the intervention. If there is both a clear opportunity to manipulate scores and a clear incentive—for example, in an evaluation of a home visiting model on maternal health if an assessment used to assign treatment is scored by the model developer after the cutoff is known. If there is both a clear opportunity to manipulate scores and a clear incentive, then the study does not satisfy this standard.
- Criterion B. The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) to establish the smoothness of the density of the forcing variable right around the cutoff. This is important to establish because there may be incentives for scorers to manipulate scores to make participants just eligible for the intervention group, in which case, there may be an unusual mass of participants near the cutoff. The statistical test must fail to reject the null hypothesis of continuity in the density of the forcing variable at the 5 percent significance level.
- Criterion C. The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff. There must not be strong evidence of a discontinuity at the cutoff that is obviously larger than discontinuities in the density at other points, although some small discontinuities may arise when the forcing variable is discrete.

A manuscript about an RDD study can satisfy or partially satisfy this standard if it meets the relevant criteria in Table C.2. A manuscript does not satisfy this standard if fewer than two of the three criteria are satisfied.

Table C.2. Satisfying the integrity of the forcing variable standard (standard 1)

Criterion	To completely satisfy the standard, a manuscript about the RDD study:	To partially satisfy the standard, a manuscript about the RDD study:
A. The institutional integrity of the forcing variable must be established by an adequate description of the scoring and treatment assignment process.	Must satisfy this criterion.	Must satisfy any two of the three criteria (A, B, or C).
B. The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	
C. The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	

Standard 2: Sample attrition

An RDD study must have acceptable levels of overall and differential attrition rates (see Chapter III). The samples used to calculate attrition must include all participants who were eligible to be assigned to the intervention or comparison group using the forcing variable, and not only a subset of those participants known to the researcher. For example, when the FSC score is used to assign families to a home visiting program, the assignment mechanism typically applies to all families who have an FSC assessment score, such as all postpartum referrals in a geographic area who have high risk scores (based on pre-screening) that make them eligible for in-depth FSC assessment. An RDD study that examines the impact of a home visiting program using a risk assessment score as the assignment variable could have acceptable levels of attrition only if it can identify the full set of families who have satisfied risk assessment requirements and have a risk assessment score. Put another way, attrition cannot be assessed unless all participants who were eligible to be assigned to conditions are known and for all of these participants, their assigned condition must be known.

However, attrition can be assessed within exogenous subgroups, meaning a subgroup identified using a variable that is exogenous to intervention participation. For example, attrition could be assessed separately within each site. Also, attrition can be calculated within a bandwidth around the cutoff value of the forcing variable. Attrition needs to be assessed separately for each contrast of interest.

The way that attrition rates are calculated determines whether a manuscript about an RDD study satisfies this standard completely or partially. Criterion A lists approaches that must be used for a manuscript to completely satisfy this standard. Criterion B lists other approaches that may be used but only allow a manuscript to partially satisfy this standard. Whereas the approaches in criterion A require the author to either use approved methods for statistically adjusting for the forcing variable or apply an acceptable bandwidth for values of the forcing variable, the approaches in criterion B may not provide as accurate an adjustment for the forcing variable. As a result, the approaches in criterion B could result in measures of overall and differential attrition at the cutoff that are less accurate.

- **Criterion A.** ***The reported combination of overall and differential attrition rates must be shown to be low using at least one of the following approaches, which have the potential to adjust for the forcing variable most accurately:***
  - Authors must report the predicted mean attrition rate at the cutoff estimated using data from below the cutoff and the predicted mean attrition rate at the cutoff estimated using data from above the cutoff. Both numbers must be estimated using a statistical model that controls for the forcing variable using the same approach that was used to estimate the impact on the outcome. Specifically, the impact on attrition must be estimated either (A) using exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (B) using the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. For the purpose of applying this standard, the overall attrition rate will be defined as the average of the predicted mean attrition rates on either side of the cutoff, and the differential attrition rate will be defined as the difference in the predicted mean attrition rates on either side of the cutoff.
  - Authors must calculate overall and differential attrition for the sample inside the bandwidth used for the impact analysis, with or without adjusting for the forcing variable. Although authors do not need to adjust for the forcing variable using this approach, other than by applying the bandwidth, the value of the forcing variable must be known for all participants so that the bandwidth can be applied.
- **Criterion B.** ***The reported combination of overall and differential attrition rates must be shown to be low when calculated using one of the following approaches, which may not provide as accurate an adjustment for the forcing variable as one of the two approaches outlined under criterion A.***
  - Authors can calculate overall and differential attrition for the entire research sample, adjusting for the forcing variable.
  - Authors can calculate overall and differential attrition for the entire research sample without adjusting for the forcing variable.

If authors calculate overall and differential attrition both ways—that is, both with and without adjusting for the forcing variable—then HomVEE will review both and assign the highest possible rating to this part of the study design. Note that approaches should not be mixed; that is, if the rating is based on an overall attrition rate calculated without an adjustment for the forcing variable, then the differential attrition rate should also be unadjusted. Unlike the approaches in Criterion A, it is possible to assess attrition using the full research sample even when the value of the forcing variable is unknown for some participants, as long as the assigned conditions of all participants is known.

A manuscript about an RDD study can satisfy or partially satisfy this standard if it meets the relevant criteria in Table C.3. A manuscript does not satisfy this standard if attrition information is not available or if neither of the criteria in the table are met.

Table C.3. Satisfying the attrition standard (standard 2)

Criterion	To completely satisfy the standard, a manuscript about the RDD study:	To partially satisfy the standard, a manuscript about the RDD study:
A. The reported combination of overall and differential attrition rates is low using an approach among those that have the potential to most accurately adjust for the forcing variable.	Must satisfy this criterion.	Does not need to satisfy this criterion.
B. The reported combination of overall and differential attrition rates is low when calculated using an approach among those that may not provide as accurate an adjustment for the forcing variable.	Does not need to satisfy this criterion.	Must satisfy this criterion.

Standard 3: Continuity of the relationship between the outcome and the forcing variable

To obtain a consistent impact estimate using an RDD, there must be evidence that in the absence of the intervention, there would be a smooth relationship between the outcome and the forcing variable at the cutoff score. This condition is needed to ensure that any observed discontinuity in the outcomes of intervention and comparison group participants at the cutoff can be attributed to the intervention.

This smoothness condition cannot be checked directly, although two indirect approaches could be used. The first approach is to test whether, conditional on the forcing variable, key *baseline* covariates that are correlated with the outcome variable (that is, race/ethnicity, socioeconomic status, and any measures of the outcome that were assessable at baseline) are continuous at the cutoff. This means that the intervention must have no impact on baseline covariates at the cutoff. Particularly important baseline covariates for this analysis are preintervention measures of the key outcome variables.

The second approach for assessing the smoothness condition is to use statistical tests or graphical analyses to examine whether there are discontinuities in the outcome-forcing variable relationship at values away from the cutoff. This process involves testing for impacts at values of the forcing variable where there should be no impacts, such as the medians of points above or below the cutoff value (Imbens & Lemieux, 2008). The presence of such discontinuities would imply that the relationship between the outcome and the forcing variable at the cutoff may not be truly continuous, suggesting that observed impacts at the cutoff may not be due to the intervention.

Three criteria determine whether a manuscript about an RDD study satisfies this standard.

- **Criterion A.** ***Baseline equivalence on key covariates, as identified in the review protocol, must be established at the cutoff value of the forcing variable.*** This involves calculating an impact at the cutoff on the covariate of interest, and the study must either (1) use exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (2) use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. Authors may exclude sample members from this analysis for reasons that are clearly exogenous to intervention participation. For example, authors may calculate baseline equivalence using only data within the bandwidth that was used to estimate the impact on the outcome. The burden of proof falls on the authors to demonstrate that any sample exclusions were made for exogenous reasons.

The baseline equivalence standards applicable to RCT and NED studies also apply to the results from this analysis; see Chapter III. Specifically, if the impact for any covariate is greater than 0.25 standard deviation in absolute value, based on the variation of that characteristic in the pooled sample, this criterion is not satisfied. If the impact for a covariate is between 0.05 standard deviation and 0.25 standard deviation, the statistical model used to estimate the average treatment effect on the outcome must include a statistical adjustment for that covariate to satisfy this criterion. Differences of less than or equal to 0.05 require no statistical adjustment.

For dichotomous covariates, authors must provide the predicted mean covariate value— that is, the predicted probability—at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff.

Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the covariate at the cutoff. These predicted probabilities are needed so that HomVEE reviewers can transform the impact estimate into standard deviation units.

If the attrition standard is at least partially satisfied, then the equivalence criterion can be demonstrated using data not in the analytic sample, such as data from a different year, cohort, or site. However, all other requirements specified above apply, including using an acceptable bandwidth and/or functional form, and excluding sample members only for clearly exogenous reasons. The review leadership team, in consultation with content experts, has discretion to determine that the sample is too different from the context in the study sample to satisfy this criterion.

If the attrition standard is not met, this analysis must be conducted using only participants with nonmissing values of the key outcome variable used in the manuscript. Exogenous exclusions from that sample are allowed. For example, participants outside of an acceptable bandwidth can be excluded.

- **Criterion B.** ***There must be no evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.*** An example of a "satisfactory explanation" is that the discontinuity corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value. Another example could be a known structural property of the assignment variable, for example, if the assignment variable is a construct involving the aggregation of both continuous and discrete components. The graphical analysis— such as a scatter plot of the outcome and forcing variable using either the raw data or averaged/aggregated data within bins/intervals—must not show a discontinuity at any forcing variable value within the bandwidth (or, for the full sample if no bandwidth is used) that is larger than two times the standard error of the impact estimated at the cutoff value, unless a satisfactory explanation of that discontinuity is provided. (The standard error at the cutoff value is used because authors may not report the standard error at the point of the observed discontinuity.)
- **Criterion C.** ***There must be no evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.*** The statistical tests must use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome and be conducted for at least four values of the forcing variable below the cutoff and four values above the cutoff; these values can be either within or outside the bandwidth. At least 95 percent of the estimated impacts on the outcome at other values of the forcing variable must be

statistically insignificant at the 5 percent significance level. For example, if impacts are estimated for 20 values of the forcing variable, then at least 19 of them must be statistically insignificant.[71]

A manuscript about and RDD study can satisfy or partially satisfy this standard if it meets the relevant criteria in Table C.4. A manuscript does not satisfy this standard if criterion A is not satisfied, or if both criteria B and C are not satisfied.

Table C.4. Satisfying the continuity of the relationship between the outcome and the forcing variable standard (standard 3)

Criterion		To completely satisfy the standard, a manuscript about the RDD study:	To partially satisfy the standard, a manuscript about the RDD study:
A.	Baseline equivalence on key covariates	Must satisfy this criterion.	Must satisfy this criterion.
B.	No evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value	Must satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C.	No evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value	Must satisfy this criterion.	

Standard 4: Functional form and bandwidth

Unlike with RCTs, statistical modeling plays a central role in estimating impacts in an RDD study. The most critical aspects of the statistical modeling are the functional form specification of the relationship between the outcome variable and the forcing variable and the appropriate range of forcing variable values used to select the analysis sample, that is, the *bandwidth* around the cutoff value. Six criteria determine whether a manuscript about an RDD study satisfies this standard.

- **Criterion A.** ***The local average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable.*** For both bias and variance considerations, it is never acceptable to estimate an impact by comparing the mean outcomes of intervention and comparison group members without adjusting for the forcing variable (even if there is a weak relationship between the outcome and forcing variable).
- **Criterion B.** ***The authors should use a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.*** For example, a bandwidth selection procedure described in an article published in a peer- reviewed journal that describes the procedure and demonstrates its effectiveness would be a justified bandwidth. An article published in an applied journal where the procedure happens to be used does not count as justification. A manuscript about a study that does not use a justified bandwidth does not completely satisfy this standard but could partially satisfy this standard if criterion C is satisfied.

[71] If impacts are estimated for fewer than 20 values of the forcing variable, all of them must be statistically insignificant at the 5 percent significance level.

- **Criterion C.** ***If the authors do not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a "best fit" regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.*** For an impact estimate to meet this criterion, the functional form of the relationship between the outcome and forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R-squared.
- **Criterion D.** ***The manuscript needs to provide evidence that the findings are robust to varying bandwidth or functional form choices.*** At least one of five types of evidence is sufficient to meet this criterion:[72]
  - In the case that criterion B applies, the sign and significance of impact estimates must be the same for a total of at least two different justified bandwidths. For example, this criterion would be satisfied if the sign and significance of an impact are the same using a bandwidth selected by cross-validation[73] and a bandwidth selected by the method described in Imbens and Kalyanaraman (2012). Two impact estimates are considered to have the same significance if they are both statistically significant at the 5 percent significance level, or if neither of them is statistically significant at the 5 percent significance level. Two impact estimates are considered to have the same sign if they are both positive, both negative, or if one is positive and one is negative, but neither are statistically significant at the 5 percent significance level.
  - In the case that criterion B applies, the sign and significance of impact estimates must be the same for at least one justified bandwidth and at least two additional bandwidths that are not justified.
  - In the case that criterion C applies, the sign and significance of impact estimates must be the same using a total of at least two different goodness-of-fit measures to select functional form. For example, this criterion would be satisfied if the impact corresponding to the functional form selected using the Akaike Information Criterion is the same sign and significance as an impact corresponding to the functional form selected using the regression *R*-squared. Note that both measures may select the same functional form.
  - In the case that criterion C applies, the sign and significance of impact estimates must be the same for at least three different functional forms, including the "best fit" regression.
  - If the manuscript meets both criteria B and C, then the sign and significance of impact estimates must be the same for the impact estimated within a justified bandwidth and the impact estimated using a "best fit" regression.
- **Criterion E.** ***The manuscript must include a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.*** The display cannot be obviously inconsistent with the choice of bandwidth and the functional form specification for the analysis. Specifically, if the authors use a particular functional form for the outcome-forcing variable relationship, then the manuscript must show graphically that this functional form fits the scatter plot

[72] If a manuscript about a study presents more than one type of evidence, and one type shows findings are robust while another type does not, then this criterion is still satisfied. That is, manuscript ratings are not penalized when authors conduct more sensitivity analyses.
[73] An implementation of cross-validation for RDD analysis is described by Imbens and Lemieux (2008).

reasonably well, and if the authors use a local linear regression, then the scatter plot must show that the outcome-forcing variable relationship is indeed reasonably linear within the chosen bandwidth.

- **Criterion F.** ***The relationship between the forcing variable and the outcome must not be constrained to be the same on both sides of the cutoff.***

A manuscript about a study can satisfy or partially satisfy this standard if it meets the relevant criteria in Table C.5. A manuscript does not satisfy this standard if either criterion A or criterion E is not satisfied or if both criteria B and C are not satisfied.

Table C.5. Satisfying the functional form and bandwidth standard (standard 4)

Criterion		To completely satisfy the standard, a manuscript about the RDD study:	To partially satisfy the standard, a manuscript about the RDD study:
A.	The local average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable.	Must satisfy this criterion.	Must satisfy this criterion.
B.	The authors should use a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.	Must satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C.	If the authors do not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then they may estimate impacts using a "best fit" regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.	Does not need to satisfy this criterion.	
D.	The manuscript needs to provide evidence that the findings are robust to varying bandwidth or functional form choices.	Must satisfy this criterion.	Does not need to satisfy this criterion.
E.	The manuscript must include a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.	Must satisfy this criterion.	Must satisfy this criterion.
F.	The relationship between the forcing variable and the outcome must not be constrained to be the same on both sides of the cutoff.	Must satisfy this criterion.	Does not need to satisfy this criterion.

Standard 5: Fuzzy regression discontinuity design

In a *sharp* RDD, all intervention group members receive intervention services and no comparison group members receive services. In a fuzzy regression discontinuity design (FRDD), some intervention group members do not receive intervention services or some comparison group members do receive intervention services, but there is still a substantial discontinuity in the probability of receiving services at the cutoff. In an FRDD analysis, the impact of service receipt is calculated as a ratio. The numerator of the ratio is the RDD impact on an outcome of interest. The denominator is the RDD impact on the probability of receiving services. This analysis is typically conducted using either two-stage least squares (2SLS) or a

Wald estimator. FRDD analysis is analogous to a complier average causal effect (CACE) or local average treatment effect analysis—consequently many aspects of this standard are analogous to the WWC standards for CACE analysis in the context of RCTs, which HomVEE also applies (see Chapter III, Section A.3.c. about treatment on the treated analyses).

The internal validity of an FRDD estimate depends primarily on three conditions. The first condition, known as the exclusion restriction, requires that the only channel through which assignment to the intervention or comparison groups can influence outcomes is by affecting take-up of the intervention being studied (Angrist et al., 1996). When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may be attributable to other factors differing between the intervention and comparison groups. The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels.

However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.

The second condition for the internal validity of an FRDD estimate is that the discontinuity in the probability of receiving services at the cutoff needs to be large enough to limit the influence of finite sample bias. The FRDD scenario can be interpreted as an instrumental variables (IV) model in which falling above or below the cutoff is an instrument for receiving intervention services (the participation indicator). IV estimators will be subject to finite sample bias if there is not a substantial difference in service receipt on either side of the cutoff, that is, if the instrument is "weak" (Stock & Yogo, 2005). FRDD impacts need not be estimated using 2SLS methods—for example, they can be estimated using Wald estimators—but authors must run the first-stage regression of the participation indicator on the forcing variable and the indicator for being above or below the cutoff, and provide either the *F* statistic or the *t* statistic from this regression.

The third condition for the internal validity of an FRDD estimate is that two relationships need to be modeled appropriately: the relationship between the forcing variable and the outcome of interest (standard 4) and the relationship between the forcing variable and receipt of services.

Ideally, the FRDD impact would be estimated using a justified bandwidth and functional form, where justification is focused on the overall FRDD impact, not just the numerator or denominator separately. Several methods have been discussed in the literature for selecting a justified bandwidth that targets the ratio (such as Calonico, Cattaneo, & Titiunik, 2014; Imbens & Kalyanaraman, 2012). However, in practice authors often use the bandwidth for the numerator of the FRDD, which is consistent with advice from Imbens and Kalyanaraman (2012).[74]

Eight criteria determine whether a manuscript about an RDD study satisfies this standard. All eight criteria are waived for impact estimates calculated using a reduced form model (in which the outcome is modeled as a function of the forcing variable, an indicator for being above or below the cutoff, and

[74] Imbens and Kalyanaraman (2012, p. 14) wrote, "In practice, this often leads to bandwidth choices similar to those based on the optimal bandwidth for estimation of only the numerator of the RD estimate. One may therefore simply wish to use the basic algorithm ignoring the fact that the regression discontinuity design is fuzzy."

possibly other covariates, but the *participation* indicator is not included in the model). This type of model is analogous to an ITT analysis in the context of RCTs.[75]

- **Criterion A.** ***The participation indicator must be a binary indicator for taking up at least a portion of the intervention.*** For example, the participation indicator could be a binary indicator for receiving any positive dosage of the intervention.
- **Criterion B.** ***The estimation model must have exactly one participation indicator.***
- **Criterion C.** ***The indicator for being above or below the cutoff must be a binary indicator for the intervention and comparison groups to which participants are assigned.***
- **Criterion D.** ***The same covariates, one of which must be the forcing variable, must be included in the analysis that estimates the impact on participation and the analysis that estimates the impact on outcomes.*** In the case of 2SLS estimation, this means that the same covariates must be used in the first and second stages.
- **Criterion E.** ***The FRDD estimate must have no clear violations of the exclusion restriction.*** Defining participation inconsistently between the assigned intervention and assigned comparison groups would constitute a clear violation of the exclusion restriction. Therefore, authors must report a definition of take-up that is the same across assigned groups. Another violation of the exclusion restriction is the scenario in which assignment to the intervention group changes the behavior of participants even if they do not take up the intervention itself. In this case, the treatment assignment might have effects on outcomes though channels other than the take-up rate. There must be no clear evidence that assignment to the intervention influenced the outcomes of participants through channels other than take-up of the intervention.
- **Criterion F.** ***The manuscript must provide evidence that the forcing variable is a strong predictor of participation in the intervention.*** In a regression of program participation on a treatment indicator and other covariates, the coefficient on the treatment indicator must report a minimum *F* statistic of 16 or a minimum *t* statistic of 4.[76] For FRDD studies with more than one indicator for being above or below the cutoff, see the WWC Version 4.1 Standards for RCTs that report CACE estimates for the minimum required first-stage *F* statistic.
- **Criterion G.** ***Authors must use a local regression or related nonparametric approach in which FRDD impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature.*** Ideally, this

[75] An important consideration when interpreting and applying these standards is that they are focused on the causal validity of impact estimates, not on appropriate interpretation of impact estimates. While the reduced form impact estimate may be a valid estimate of the effect of being below (or above) the RDD cutoff, interpreting that impact can be challenging in some contexts. In particular, while the reduced form RDD impact is methodologically analogous to the intent to treat (ITT) impact from an RCT, the substantive interpretation can be entirely different. Addressing these interpretive issues is beyond the scope of these standards, but we urge users of these standards to think carefully about interpretation.

[76] Stock and Yogo (2005). The *F* statistic must be for the instrument only—*not* the *F* statistic for the entire first stage regression. If the unit of assignment does not equal the unit of analysis, then the *F* statistic or *t* statistic must account for clustering using an appropriate method (such as boot-strapping, hierarchical linear modeling [HLM], or the method proposed by Lee and Card, 2008). Also, in a working paper, Fier, Lemieux, and Marmer (2016) suggested that in the FRDD context, the minimum first-stage *F* statistic that ensures asymptotic validity of a 5 percent two-sided test is much higher than would be required in a simple IV setting; specifically, they suggest 135. Until a published paper provides an *F* statistic cutoff that is appropriate for FRDD studies that use a justified bandwidth, the *F* statistic of 16 will be used as the interim criterion for assessing instrument strength.

method would be justified for the FRDD impact estimate, not just the numerator of the FRDD estimate. However, two other approaches are acceptable. First, it is acceptable to use separate bandwidths for the numerator and denominator, if both are selected using a justified approach, such as the IK algorithm applied separately to the numerator and denominator. Second, it is acceptable to use the bandwidth selected for the numerator if that bandwidth is smaller than or equal to a justified bandwidth selected for the denominator.

- **Criterion H.** ***If Criterion G is not met, the manuscript can still partially satisfy the standard if the FRDD impact is estimated using a bandwidth that is only justified for the numerator, even if it is larger than a bandwidth justified for the denominator.*** This criterion is also satisfied if the denominator is estimated using a "best fit" functional form. That is, the functional form of the relationship between program receipt and the forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R-squared.

A manuscript about an RDD study can satisfy or partially satisfy this standard if it meets the relevant criteria in Table C.6. A manuscript does not satisfy this standard if any of criteria A–F are not satisfied, or if both criteria G and H are not satisfied.

Table C.6. Satisfying the fuzzy regression discontinuity design standard (standard 5)

Criterion	To completely satisfy the standard, the manuscript about the RDD study:	To partially satisfy the standard, the manuscript about the RDD study:
A. The participation indicator must be a binary indicator	Must satisfy this criterion.	Must satisfy this criterion.
B. The estimation model must have exactly one participation indicator	Must satisfy this criterion.	Must satisfy this criterion.
C. The indicator for being above or below the cutoff must be a binary indicator for the groups	Must satisfy this criterion.	Must satisfy this criterion.
D. The same covariates must be included in (1) the analysis that estimates the impact on participation and (2) the analysis that estimates the impact on outcomes	Must satisfy this criterion.	Must satisfy this criterion.
E. No clear violations of the exclusion restriction	Must satisfy this criterion.	Must satisfy this criterion.
F. Evidence that the forcing variable is a strong predictor of participation in the intervention	Must satisfy this criterion.	Must satisfy this criterion.
G. Local regression or related nonparametric approach with a justified bandwidth	Must satisfy this criterion.	Does not need to satisfy this criterion.
H. Local regression or related nonparametric approach with a bandwidth that is only justified for the numerator or the denominator is estimated using a best fit functional form	Does not need to satisfy this criterion.	Must satisfy this criterion.

D. Applying standards to studies that report multiple impact estimates

Some manuscripts about RDD studies report multiple separate impacts (findings), for example, impacts for different outcomes or subgroups of interest. Each of the standards described above will be applied to each outcome-subgroup combination, resulting in a separate rating for each combination. The overall rating for the manuscript will be the highest rating attained by any outcome-subgroup combination that is eligible for review by HomVEE and will apply to only the combination(s) with that rating. In Section E, we address the special case of impacts that are pooled or aggregated across multiple combinations of forcing variables, cutoffs, and samples.

E. Applying standards to studies that involve aggregate or pooled impacts

Some manuscripts about RDD studies may report pooled or aggregate impacts for some combinations of forcing variables, cutoffs, and samples. By "pooled impact," we mean that data from each combination of forcing variable, cutoff, and sample are standardized and grouped into a single dataset for which a single impact is calculated. By "aggregate impact," we mean a weighted average of impacts that are calculated separately for every combination of forcing variable, cutoff, and sample.

The overall rating for the manuscript will be the highest rated impact—including pooled and aggregate impacts—presented in the manuscript. Authors may improve the rating of a pooled or aggregate impact by excluding combinations of forcing variables, cutoffs, and samples rate low for reasons that are clearly exogenous to intervention participation. For example, in a multisite study, a site that fails the institutional check for manipulation could be excluded from the aggregate impact, resulting in a higher rating for the aggregate impact. However, potentially endogenous exclusions—those potentially influenced by the intervention—will not improve the rating of an aggregate impact because standards will be applied as if

those exclusions were not made. For example, excluding sites that have a high differential attrition rate from an aggregate impact will not improve the rating of that impact because for the purpose of applying the attrition standard, we will include those sites. The burden of proof falls on the authors to demonstrate that any exclusions from the aggregate impact were made for exogenous reasons.

For each impact that is based on a single forcing variable, cutoff, and sample, the standards can be directly applied as stated in Section C.

For pooled or aggregate impacts that are based on multiple forcing variables, cutoffs, or samples, additional guidance for applying the standards is provided next.

Standard 1: Integrity of the forcing variable

- **Criterion A.** ***If the institutional integrity of the forcing variable is not satisfied for any combination of forcing variable, cutoff, and sample that are included in a pooled or aggregate impact, then this criterion is not satisfied for that pooled or aggregate impact.*** However, it is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion. For example, if a pooled or aggregate impact is estimated using data from five sites, and the institutional integrity of the forcing variable is not satisfied in one of those five sites, then the pooled or aggregate impact does not satisfy this criterion. However, a pooled or aggregate impact estimated using data from only the four sites for which the institutional integrity of the forcing variable is satisfied would satisfy this criterion.
- **Criterion B.** ***For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.*** In the case of a pooled impact, applying an appropriate statistical test to the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.
- **Criterion C.** ***For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.*** In the case of a pooled impact, providing a single figure based on the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

Standard 2: Attrition

In the case of a pooled impact, the attrition standard described in Section C can be applied directly if the authors calculate and report overall and differential attrition using the pooled sample. Any sample excluded from calculating the pooled or aggregate impact for reasons of endogeneity—that is, because the sample was potentially influenced by the intervention—cannot be excluded from the attrition calculation.

In the case of an aggregate impact, the attrition standard can be applied to the overall and differential attrition rates calculated as weighted averages of the overall and differential rates calculated for each unique combination of forcing variable, cutoff, and sample that contribute to the aggregate impact. Authors must calculate overall and differential attrition for each of those unique combinations in a way that is consistent with the standard described in Section C, and the weights used in aggregation must be the same weights used to calculate the weighted impact being reviewed. The attrition standard described in Section C is then applied to the combination of overall and differential attrition based on the weighted average.

Standard 3: Continuity of the relationship between the outcome and the forcing variable

- **Criterion A.** ***In the case of a pooled impact, this criterion can be applied as described in Section C without modification.*** In the case of an aggregate impact, baseline equivalence can be established by applying the same aggregation approach to the impacts on baseline covariates as is used to aggregate impacts on outcomes.
- **Criterion B.** ***In the case of a pooled impact, this criterion can be applied as described in Section C without modification.*** In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. Specifically, there must not be evidence of a discontinuity larger than twice the standard error of the impact at any noncutoff value within the bandwidth of any forcing variable for any sample. This means that a graphical analysis must be presented for every combination of forcing variable, cutoff, and sample. In cases where impacts from disjointed—that is, nonoverlapping—samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion, such an exclusion is considered exogenous.
- **Criterion C.** ***In the case of a pooled impact, this criterion can be applied as described in Section C without modification.*** In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. That is, at least 95 percent of estimated impacts at values of the forcing variables other than the cutoffs, across all samples, must be statistically insignificant. In cases where impacts from disjointed samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion; such an exclusion is considered exogenous.

Standard 4: Functional form and bandwidth

In the case of a pooled impact, this standard can be applied as described in Section C without modification.

In the case of an aggregate impact, criteria A, B, C, E, and F of this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate because they do not satisfy one of those criteria will be treated as attrition. The aggregate impact will receive the lowest rating from among all of these impacts.

Criterion D can be applied only to the aggregate impact. That is, it is sufficient to demonstrate robustness of the aggregate impact—it is not necessary to show robustness of every impact included in the aggregate, although showing robustness for every individual impact is also acceptable.

Standard 5: Fuzzy regression discontinuity design

In the case of a pooled impact, this standard can be applied as described in Section C without modification.

In the case of an aggregate impact, this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate will be treated as attrition, with two exceptions—impacts may be excluded if they do not meet criterion E or F. The aggregate impact will receive the lowest rating from among all of these impacts.

F. Cluster-assignment regression discontinuity designs

Following the WWC, HomVEE considers an RDD study to be a cluster-assignment study when individuals are assigned to conditions in groups and the outcome measure is assessed for individuals within clusters. The same two screening conditions for cluster-assignment group design studies apply as are discussed in Section B of this appendix. We provide additional criteria for applying the five RDD standards to cluster-assignment RDDs here. These criteria describe how and when to use cluster- or individual-level data to satisfy each RDD standard.

As with cluster group design studies, cluster RDDs can satisfy HomVEE standards for effects of an intervention on individuals or on clusters. HomVEE initially reviews a manuscript about a cluster RDD study for evidence of an intervention's effect on individuals. If an effect on individuals cannot be credibly demonstrated, then HomVEE reviews the evidence of an intervention's effect on clusters, where changes in the composition of individuals within the clusters may influence the observed effect. When a manuscript about an RDD study satisfies the standards for effects of the intervention on individuals, it may be eligible for the rating of high.

However, the observed impact estimate in an RDD manuscript that satisfies standards for effects on clusters but not on individuals potentially represents a combination of the effect of the intervention on individuals and a composition effect due to different types of individuals entering intervention and comparison clusters. Therefore, when an RDD manuscript satisfies only those standards for effects on clusters, it is only eligible to be rated moderate.

Standards 1, 4, and 5

These standards are assessed in the same way whether the manuscript is being reviewed for evidence of an intervention's effect on individuals or on clusters. Each of these standards is assessed using the criteria described in Section C, using individual-level or cluster-level data. For example, if neighborhoods are assigned to conditions and the authors estimate the impact of a home visiting program on family economic self-sufficiency using family poverty index scores averaged to the neighborhood level, then criteria B and C of Standard 1 (integrity of the forcing variable) could be assessed using neighborhood-level data or family-level data (the assessment of criterion A does not rely on study data).

Standard 2: Attrition

The attrition standard can be completely or partially satisfied in the review of a cluster RDD for effects on individuals. If the standard is not satisfied in the review for effects on individuals, then it may be partially satisfied (but not completely satisfied) in the review of the manuscript for effects on clusters.

Review of a cluster RDD for effects on individuals
In the review of a cluster RDD for evidence of effects on individuals, individuals who enter clusters after the results of assignment are known may pose a risk of bias. Therefore, the attrition standard includes an assessment of potential risk of bias from joiners. If the analytic sample includes individuals who joined clusters after random assignment and those individuals pose a risk of bias, then the attrition standard can only be partially satisfied, and the highest rating the manuscript about the study can receive is moderate.

For a manuscript about a cluster-assignment RDD study to *completely satisfy* the attrition standard in the review for evidence of effects on individuals, the manuscript must meet the following three requirements:

- Limit the risk of bias from individuals who entered clusters after assignment as described in WWC Version 4.1 standards for cluster RCTs (see Chapter III).
- Meet the same requirements for completely satisfying the standard using individual-level data within nonattriting clusters, applying an acceptable reference sample as the denominator of the attrition calculation (see Chapter III).
- Meet the requirements for completely satisfying the standard as described in C of this RDD standard by using cluster-level data.

To *partially satisfy* the standard in the review for evidence of effects on individuals, the manuscript must meet the following requirements:

- Limit the risk of bias from individuals who entered clusters after assignment.
- Meet the same requirements for completely or partially satisfying the standard using individual-level data within nonattriting clusters, applying an acceptable reference sample.
- Meet the requirements for completely or partially satisfying the standard as described in Section C of this RDD standard by using cluster-level data.

Review of a cluster RDD for effects on clusters

In the review of a cluster RDD for evidence of effects on clusters, the manuscript cannot *completely satisfy* the attrition standard because of the risk that impact estimates may in part reflect compositional changes.

To *partially satisfy* the standard in the review of evidence of effects on clusters, the manuscript must meet the following two requirements:

- Meet the requirements for completely or partially satisfying the standard as described in Section C by using cluster-level data.
- Demonstrate that the analytic sample of individuals used to estimate the impact of the intervention is representative of the clusters as described in HomVEE standards for cluster RCTs (see handbook Chapter III, Section C.1). The attrition calculations for this representativeness requirement must be performed using an approach that would completely or partially satisfy the RDD attrition standard described in Section C above.

Standard 3: Continuity

The continuity standard can be completely or partially satisfied in the review of a cluster RDD for effects on individuals. If the standard is not satisfied in the review for effects on individuals, then it may be partially satisfied, but not completely satisfied, in the review of the manuscript for effects on clusters.

Review of a cluster RDD for effects on individuals

For a cluster RDD to *completely satisfy* this standard, the manuscript must meet the requirements for satisfying the continuity standard described in Section C, above. If the attrition standard is not satisfied in the review for effects on individuals, then criterion A of the continuity standard must be satisfied using the analytic sample of individuals—those who contribute outcome data to the impact analysis. When authors analyze outcomes aggregated to the cluster level, the analytic sample of individuals are those who

contribute outcome data to the cluster-level averages. These requirements can be met using individual-level or cluster-level data.

For a cluster RDD to *partially satisfy* the standard, the manuscript must meet the requirements for partially satisfying the continuity standard described in Section C. Again, if the attrition standard is not satisfied in the review for effects on individuals, then criterion A of the continuity standard must be satisfied using the analytic sample of individuals.

Review of a cluster RDD for effects on clusters

In the review of a cluster RDD for evidence of effects on clusters, the manuscript cannot *completely satisfy* the continuity standard because of the risk that impact estimates may in part reflect compositional changes.

To *partially satisfy* the standard in the review of evidence of effects on clusters, the manuscript about the study must meet the following requirements:

- Meet the requirements for completely or partially satisfying the continuity standard as described in Section C, where criterion A of the standard must be satisfied using the analytic sample of clusters if the attrition standard is not satisfied in the review for effects on clusters.
- Demonstrate that the sample of individuals used to assess criterion A of the continuity standard is representative of the clusters as described in HomVEE standards for cluster RCTs (see handbook Chapter III, Section C.1).
- Demonstrate that the samples of individuals used to assess criteria B and C of the continuity standard and the analytic sample used to estimate impacts are representative of the clusters as described in WWC Version 4.1 standards for cluster RCTs. Frequently, the samples used to assess these criteria will be identical to those used to assess impacts, so this representativeness requirement need only be assessed once.

G. Reporting requirement for studies with clustered sample

As is the case in RCTs, clusters of individuals or families might be assigned in groups to the intervention and comparison conditions. Clustering affects standard errors but does not lead to biased impact estimates, so if authors do not appropriately account for the clustering of students, a manuscript about an RDD study can still rate moderate or high if it satisfies the standards described above. However, because the statistical significance of findings is used for the rating of the effectiveness of an intervention, when observations are clustered into groups and the unit of assignment, the cluster, differs from the unit of analysis, the individual, authors must account for clustering using an appropriate method in order for findings reported by the author to be included in the rating of effectiveness. Appropriate methods including boot-strapping, multilevel linear modeling, or the method proposed by Lee and Card (2008). If the authors do not account for clustering, then HomVEE will not rely on the statistical significance of the findings from the manuscript.

H. Reporting requirement for dichotomous outcomes

For dichotomous outcomes, authors must provide the predicted mean outcome—that is, the predicted probability—at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff. Both predicted probabilities must be calculated using

the same statistical model that is used to estimate the impact on the outcome at the cutoff. These predicted probabilities are needed in order for findings reported by the author for those outcomes to be included in the rating of effectiveness.

This page has been left blank for double-sided copying.

References

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

Fier, D., Lemieux, T., & Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business and Economic Statistics, 34*(2), 185–196.

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies, 79*(3), 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615–635.

Lee, D., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*(2), 655–674

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698–714.

Reardon, S., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating- score variables. *Journal of Research on Educational Effectiveness*, *5*(1), 83–104.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In J. Stock & D. W. K. Andrews (Eds.), *Identification and inference for econometric models: Essays in* Honor *of Thomas J. Rothenberg* (pp. 80–108). Cambridge, MA: Cambridge University Press.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141.

This page has been left blank for double-sided copying.

Appendix D

Standards and Reporting Procedures for Single-Case Design Research

This page has been left blank for double sided copying.

This appendix replicates the What Works Clearinghouse Version 4.1 standards and procedures for research with this design, except for minor wording changes to tailor them to the HomVEE context.[77]

A. Identifying whether a manuscript is about a single-case design study

These standards are intended to guide reviewers in identifying and evaluating single-case design research (SCDs). If a study is an eligible SCD, any manuscript about it is reviewed using the rating criteria to determine whether it receives a rating of high, moderate, or low.[78]

Eligible SCDs are identified by the following features:

- An individual **case** is the unit of intervention administration and data analysis. A case may be a single participant or a cluster of participants, such as a home visiting clients in a given county or ZIP code.
- Within the design, the case can provide its own control for purposes of comparison. For example, the case's series of outcome variables prior to the intervention is compared with the series of outcome variables during and after receiving the intervention.
- The outcome variable is measured *repeatedly* within and across *different* conditions or levels of the independent variable. These different conditions are referred to as **phases**, such as the first baseline phase, first intervention phase, second baseline phase, and second intervention phase.

The standards for SCDs apply to a wide range of designs, including ABAB designs, multiple baseline designs, alternating and simultaneous intervention designs, changing criterion designs, and variations of these core designs like multiple probe designs. Even though SCDs can be augmented by including one or more independent comparison cases, in this document, these SCD standards address only the core SCDs and are not applicable to the augmented independent comparison SCDs.

B. Determining a manuscript rating

If the study appears to be an SCD, the following rules are used to determine whether the manuscript about the study rates high, moderate, or low. In order to meet standards, the following design criteria must be present, as illustrated in Exhibit D.1:

1. Data availability

- Authors of manuscripts about SCD studies must provide raw data in graphical or tabular format to permit visual analysis of the data to help HomVEE assess whether the study meets requirements for internal validity for SCDs.

2. Independent variable

- The independent variable indicating assignment to the intervention must be systematically manipulated; the researcher will determine when and how the independent variable conditions change.

[77] What Works Clearinghouse. (2020). Handbooks and Other Resources: Procedures and Standards Handbooks. Retrieved June 4, 2020, from https://ies.ed.gov/ncee/wwc/handbooks.

[78] For manuscripts about studies that are rated moderate or high only, HomVEE calculates a design-comparable effect size is calculated if it is possible to do so. See Section D of this appendix.

3. Inter-assessor agreement

- For each case, the outcome variable must be measured systematically over time by more than one assessor. The design needs to collect inter-assessor agreement (IAA) in each phase and at least 20 percent of the data points in each baseline and intervention condition, and the IAA must meet minimal thresholds. IAA, commonly called interobserver agreement, must be documented on the basis of a statistical measure of assessor consistency. Although there are more than 20 statistical measures to represent IAA (for example, Berk, 1979; Suen & Ary, 1989), commonly used measures include percentage or proportional agreement and Cohen's kappa coefficient, which adjusts for the expected rate of chance agreement (Hartmann, Barrios, & Wood, 2004). According to Hartmann et al., (2004), minimum acceptable values of IAA are at least 0.80, if measured by percentage agreement, and at least 0.60, if measured by Cohen's kappa. The IAA needs to meet these minimum values for each outcome *across all phases and cases*, but not separately for each case or phase. If the manuscript does not meet these minimum values for each outcome *across all phases and cases*, then it is rated low.[79]

4. Residual treatment effects (if applicable)

- Alternating treatment (AT) designs and designs with an intervening third condition are potentially subject to *residual treatment effects*—responses within phases and conditions that are caused by interventions in previous phases and conditions. When there are three or more interventions in an alternating treatment design, the reviewer must ensure that there are no residual treatment effects. If an intervention is judged to have a reasonable likelihood of residual treatment effects, the manuscript is rated low.
  - When a review team identifies an eligible alternating treatment design experiment that uses three or more interventions, the review team will ask a subject matter expert to determine whether residual treatment effects are likely given the specific interventions and outcomes in the experiment (HomVEE can rely on previous approval of similar conditions and outcomes from the subject matter expert; the plausibility of residual effects is not uniquely informed by the data in a given manuscript). HomVEE will then assign the manuscript for review and pass along the subject matter expert determination to the reviewers. Reviewers then raise any additional concerns they have about residual treatment effects as part of their reviews.
  - In most cases, the plausibility of residual treatment effects is based on theoretical and contextual considerations. Concerns about residual treatments will focus on study design and intervention characteristics, rather than on observed data.
  - If the subject matter expert and reviewer both agree that there are likely to be residual treatment effects, then the manuscript is rated low because the measures of effectiveness cannot be attributed solely to the intervention.
  - If the subject matter expert and reviewer disagree, then review team leadership will revisit the issue with the subject matter expert. If the subject matter expert and reviewer both agree that residual treatment effects are unlikely, then the reviewer will complete the review assuming there are no residual treatment effects.

[79] HomVEE will conduct author queries if the authors do not report the total percentage of sessions checked for IAA, whether IAA was checked at least once in each phase for each participant, or the IAA statistic—for example, percentage agreement—was used to demonstrate reliability. HomVEE also will conduct an author query if the authors do not specify that IAA data were collected during *each phase and for each case* for an outcome.

- Reversal-withdrawal designs, multiple baseline, and multiple-probe designs generally have longer phases than alternating treatment designs, which means more time will pass between the noncontiguous phases that will be compared (for example, between the first B and second A in an ABCAB reversal-withdrawal design); this feature may make residual effects less important even if they are present. If the reviewer and subject matter expert agree that residual effects are unlikely, or are unlikely to be meaningful, then the reviewer(s) will work with the review team leadership and subject matter experts to identify how best to proceed with the review, focusing only on the intervention of interest and the relevant comparison condition when assigning a manuscript rating (that is, ignoring any third or fourth interventions). The alternating treatment design guidance can be used as a foundation.

5. Other concerns

- **Confounding factor.** The study must not have a *confounding factor.* In SCDs, when study participants experience a different interventionist (for example, home visitor or parent manipulating the intervention condition) across baseline and intervention phases of the study, the study has a potential confounding factor. As it can sometimes be difficult to determine whether something is a confounding factor, readers are referred to the WWC Version 4.1 standards handbook for additional guidance and examples for the identification of confounding factors in SCDs.
- **Training phases, if present, cannot overlap.** Once reviewers have determined that the timing of sessions is presented consistently, they will assess concurrence and effects. In order to have concurrence, the cases still in the baseline phase must continue baseline measurement at or after the time point when a preceding case has the first intervention probe after completing their training. In other words, there can be no overlap in the training phases among the cases in the experiment.
  - If this requirement is not met, then there is no concurrence—the design cannot exclude threats to internal validity and will be rated low because there are insufficient data to evaluate the attempts to demonstrate an intervention effect.
  - If this requirement is met, the experiment can be rated moderate or high. In addition, when evaluating concurrence in multiple-probe designs, HomVEE also requires that "Each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches the prespecified intervention criterion." [80] When impacts are expected only after complete delivery of the training, the "first receives the intervention" language will be interpreted as the time point when a case has the first intervention probe after completing their training.

6. Attempts to demonstrate effect over time and data points per phase

- The manuscript must report at least three attempts to demonstrate an intervention effect at three different points in time.[81] The three demonstrations criterion is based on professional convention (Horner, Swaminathan, Sugai, & Smolkowski, 2012).

[80] If HomVEE, in consultation with subject matter experts, determines there are exceptions to this standard, these will be specified in updates to this handbook. For example, extreme child maltreatment might warrant not requiring each case requiring a probe in a session where another case first receives the intervention or reaches the prespecified intervention criterion.

[81] Although atypical, there might be circumstances in which designs without three replications meet the standards. A case must be made by the review team leadership based on input from subject matter experts, and at least two reviewers must agree with this decision.

- Depending on the design type, phases must meet criteria involving the number of data points.[82] Failure to meet any of these criteria results in a manuscript rating of low.
  - ***Reversal or withdrawal (AB).*** Must have a minimum of four phases per case with at least five data points per phase to be rated high. Must have a minimum of four phases per case with at least three data points per phase to be rated moderate. Any phases based on fewer than three data points will result in the rating of low unless otherwise determined by review team leadership.
  - ***Multiple baseline and multiple probe.*** Must have a minimum of six phases with at least five data points per phase to be rated high. Must have a minimum of six phases with at least three data points per phase to be rated moderate. Any phases based on fewer than three data points will result in the rating of low unless otherwise determined by review team leadership. The timing of the design's implementation requires a degree of concurrence when the intervention is being introduced. Otherwise, these designs cannot be distinguished from a series of separate AB designs.
  - ***Alternating treatment.*** Must have a minimum of five data points per baseline or intervention condition and at most two data points per phase to be rated high. Must have four data points per condition and at most two data points per phase to be rated moderate. Any phases based on more than two data points will result in the rating of low unless otherwise determined by review team leadership. When designs include multiple intervention comparisons—for example, A versus B, A versus C, C versus B—each intervention comparison is rated separately.
  - ***Changing criterion.*** The reversal or withdrawal (AB) design standards will be applied to changing criterion designs. Each baseline or intervention change or criterion change will be considered a phase change. As such, there should be at least three different criterion changes to establish three attempts to demonstrate an intervention effect. In some studies using this design, the researcher may reverse or change the criterion back to a prior level to further establish that the change in criterion was responsible for the outcomes observed on the dependent variable. This will be considered a phase change, as in the reversal-withdrawal design.
  - ***Multiple-probe designs.*** These designs are a special case of multiple baseline design and must meet additional criteria because baseline data points are intentionally missing.[83] Failure to meet any of these results in a manuscript rating of low.
    - Initial preintervention data collection sessions must overlap vertically. Within the first three sessions, the design must include three consecutive probe points for each case to be rated high and at least one probe point for each case to be rated moderate.
    - Probe points must be available just prior to introducing the independent variable. Within the three sessions just prior to introducing the independent variable, the design must include three consecutive probe points for each case to be rated high and at least one probe point for each case to be rated moderate.

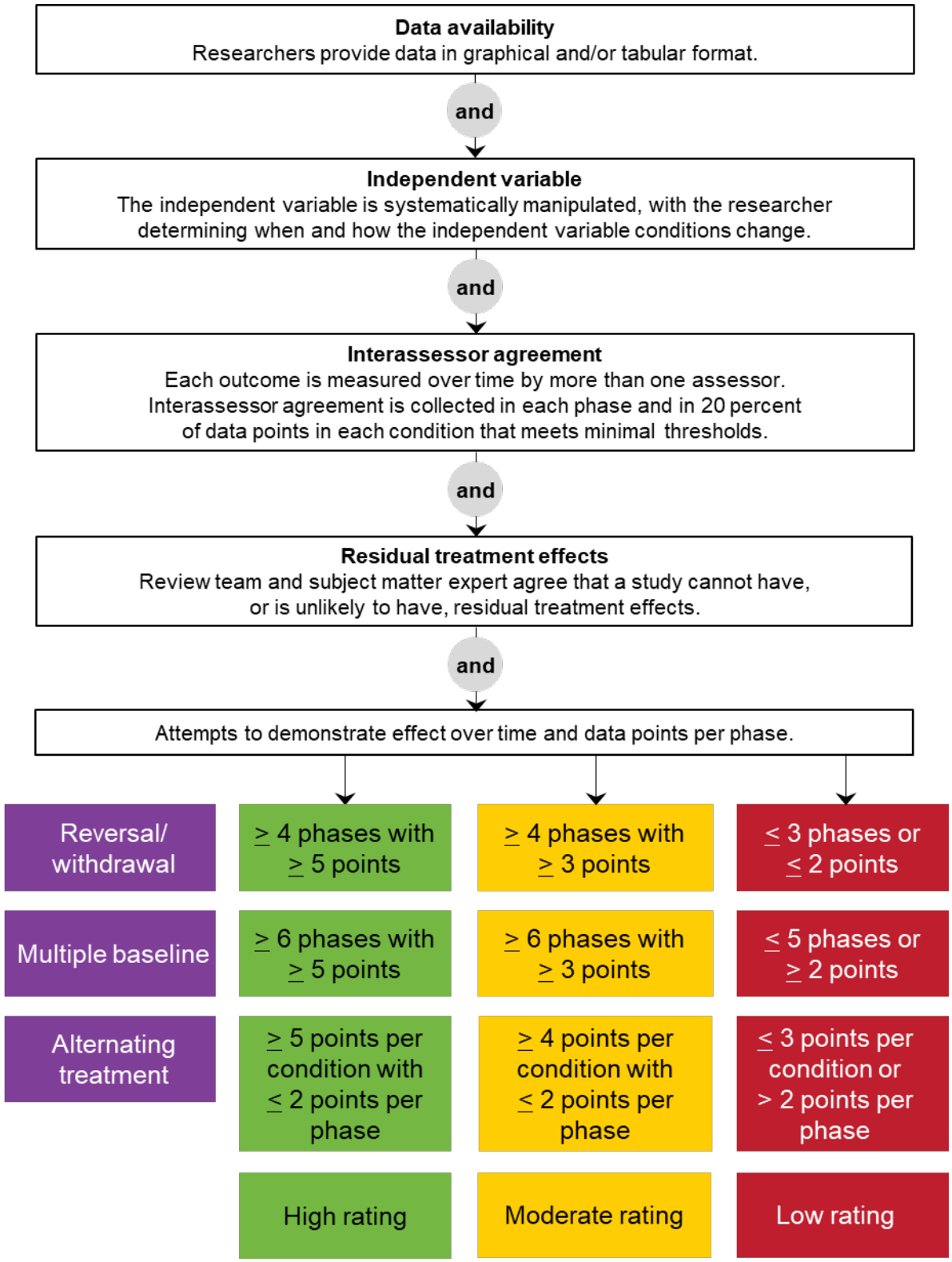
[82] If HomVEE, in consultation with subject matter experts, determines that there are exceptions to this standard, these will be specified in updates to this handbook. For example, extreme child maltreatment might warrant a lower threshold of only one or two data points.

[83] If HomVEE, in consultation with subject matter experts, determines that there are exceptions to these standards, then HomVEE will publish an update to this handbook (for example, conditions when stable data patterns necessitate collecting fewer than three consecutive probe points just prior to introducing the intervention or when collecting overlapping initial preintervention points is not possible).

    - Each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches the prespecified intervention criterion. This point must be consistent in level and trend with the case's previous baseline points.

- Reversal-withdrawal, multiple-baseline, and multiple-probe designs may have more than the minimum required number of phases required to meet standards, for example, a reversal-withdrawal design with six phases (ABABAB) or a multiple baseline design with four cases where each case has two phases.
  - The reviewer will first conduct the review considering all phases and cases (that is, review the experiment as conducted and reported). If the experiment is rated high or moderate when considering all phases and cases, then the reviewer will complete the review without separately considering subsets of phases or cases.
  - If the experiment is rated low when considering all relevant phases (for example, because some phases do not have at least three data points), the reviewer will conduct the review considering the subset of consecutive phases (in a reversal-withdrawal design) or consecutive cases (in a multiple baseline or multiple probe design) with enough points and determine whether the subset can meet standards. There may also be multiple rigorous subsets of phases. Reviewers will select the subset aimed at measuring the effectiveness of the intervention of interest. When selecting a subset of phases or cases to review, reviewers will discuss the ultimate choice review team leadership. Reviewers will document the phases and cases used in the review and the reasons why some may have been excluded from the review. This information will also be documented in HomVEE products that cite the manuscript.

Exhibit D.1. Rating determinants for single-case designs



C. Nondesign components

This handbook discusses outcome requirements and confounding factors generally in Chapter III. The nature of SCDs necessitates additional specification on elements of the two nondesign components.

1. Reliability

In SCDs, the minimum for percentage agreement—regardless of whether the metric is exact agreement or agreement within 1—is 80 percent (or .80). The minimum kappa or correlation is 0.60. IAA needs to meet these minimum values for each outcome across all phases and cases, but not separately for each case or phase. If the manuscript does not meet these minimum values for each outcome across all phases and cases, then it is rated low because the eligible outcomes do not meet requirements; more specifically, the outcomes do not meet minimum IAA thresholds.

If authors do not report that at least 20 percent of the total sessions were checked for IAA and/or that IAA was checked at least once in each phase, then the manuscript is rated low because the eligible outcomes do not meet requirements; more specifically, the outcomes do not meet minimum IAA requirements.

If authors do not report that IAA data were collected at least once for each phase or case combination, the manuscript is rated low because the eligible outcomes do not meet requirements; more specifically, the outcomes do not meet minimum IAA requirements.

When a manuscript does not report reliability statistics for an outcome measure, HomVEE will ask the authors to provide a statistic.

2. Confounding Factors:

In some SCD studies, a component of the study design or the circumstances under which the intervention was implemented are perfectly aligned, or confounded, with either the baseline or intervention phase. That is, some factor is present for only one phase and absent for other phase(s). Because it is impossible to separate the degree to which an observed effect was due to the intervention and how much was due to the confounding factor, a manuscript about a study with a confounding factor is rated low because measures of effectiveness cannot be attributed solely to the intervention.

Reviewers must decide whether there is enough information to determine that the only difference between phases that is not controlled for by design or analysis is the presence of the intervention. If not, there may be a confounding factor, and the reviewer must determine whether that factor could affect the outcome separately from the intervention. For HomVEE to determine that a confounding factor is present in the study, there must be evidence of its presence. A specific factor that is aligned with the baseline or intervention condition must be identified based on information in the manuscript or obtained from an author query.

In SCDs, home visitors or parents—collectively labeled *interventionists*—can administer the intervention to study participants. When study participants experience a different interventionist across baseline and intervention phases of the study, the study has a potential confounding factor.

As it can sometimes be difficult to determine whether something is a confounding factor, HomVEE reviewers will reference the latest WWC Version 4.1 guidelines and consult with SMEs when determining whether a potential confounding factor should affect the rating of an SCD manuscript.

D. Procedures for reporting SCD findings

For SCD studies that are rated moderate or high, HomVEE will calculate a design-comparable effect size (D-CES) where feasible and appropriate in the judgment of review team leadership. The D-CES is comparable with a standardized mean-difference effect size. This is intended to be interpreted similarly to the Hedges' *g*, the effect size HomVEE attempts to report where available for findings from RCT and NED studies (Pustejovsky, Hedges, & Shadish, 2014; Shadish, Hedges, & Pustejovsky, 2014).

1. Approach to effect sizes for SCD studies

SCDs involve multiple observations in treatment and comparison conditions for each individual. Despite the name, SCDs typically involve data from several individuals. For each individual, there are multiple observations within each treatment phase.

A D-CES can be computed for a study that has three or more participants in a design that is multiple baseline across individuals, multiple probe across individuals, or a treatment reversal $(AB)^k$ design. In each case, the numerator of the effect size is a mean of the difference between observations in the treated and comparison conditions, averaged across individuals. The denominator of the effect size is an estimate of the between-person-within-condition standard deviation. Because the observations within persons are correlated, the computation of the degrees of freedom of the denominator and the variance of the effect size is more complex than in conventional between-subjects designs. Moreover, the number of degrees of freedom in the denominator is typically close to the number of participants, which is often rather small so that the bias correction, analogous to that used to compute Hedges' *g*, is quite important.

The statistical details and formulas for computing design-comparable effect sizes are given in the next section of this appendix. For a more complete exposition, see Hedges, Pustejovsky, and Shadish (2012); Hedges, Pustejovsky, and Shadish (2013); and Pustejovsky et al. (2014).

Computing the D-CES requires access to raw outcome data by case, by observation occasion, and by treatment phase. The preferred method of obtaining raw data, if not presented in a suitable form in the manuscript being evaluated, is from the study authors. If study authors do not provide raw data but clear graphs are provided in the paper, then HomVEE reviewers may also use a graph-digitizing software to extract the individual points from a graph.

When estimating the D-CES, HomVEE reviewers will begin with the following default specifications:

1. Use restricted maximum likelihood as the default estimator.
2. Specify the intervention effect as a fixed effect.
3. Assume "no trend" at baseline or any later phases for the estimation of the D-CES in multiple baseline designs.

Review team leadership may determine, on the basis of visual analysis or an appropriate algorithm, that the underlying data do not conform to the above specifications. HomVEE may, after consultation with the content and methodological experts, either change the above specifications or not compute the D-CES, if an appropriate method is not available. HomVEE will document the rationale for any departures from the default specifications for computing the D-CES.

2. Technical details of calculating design-comparable effect sizes from single-case designs

As outlined in the section above, a D-CES can be computed for a study that has three or more participants in a design that is multiple baseline across individuals, multiple probe across individuals, or a treatment reversal design. Shadish, Hedges, and Pustejovsky (2014) provided a formula to compute the effect size $\left(d = \frac{\bar{D}}{S}\right)$ for the treatment reversal design where:

[D.1.0] $$\bar{D} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{a=1}^{k}\left(\frac{1}{n}\sum_{t=(2a-1)n+1}^{2an} Y_{ij} - \frac{1}{n}\sum_{j=(2a-2)n+1}^{(2a-1)n} Y_{ij}\right),$$

where Y_{ij} is the observation of case *i* at time *j* in phase pair a, *m* is the number of cases, *n* is the number of timepoints per phase, and *k* is the number of AB phase pairs.

[D.1.1] $$S = \sqrt{\frac{1}{2kn(m-1)}\sum_{a=1}^{2k}\ \sum_{j=(a-1)n+1=1}^{an}\ \sum_{i=1}^{m}\left(Y_{ij} - \bar{Y}_{.j}\right)^2},$$

$\bar{Y}_{.j}$ is the mean across individuals at the t^{th} time-point given by:

[D.1.2] $$\bar{Y}_{.j} = \frac{1}{m}\sum_{i=1}^{m} Y_{ij}\ .$$

The D-CES for the multiple baseline (across individuals) and multiple probe (across individuals) designs is also defined as $\left(d = \frac{\bar{D}}{S}\right)$ but where:

[D.1.3] $$\bar{D} = \frac{1}{m}\sum_{i=1}^{m}\left(\bar{Y}_{i.}^{T} - \bar{Y}_{i.}^{B}\right),$$

where $\bar{Y}_{i.}^{T}$ and $\bar{Y}_{i.}^{B}$ are the average outcomes for individual *i* within the intervention and baseline conditions, respectively, and:

[D.1.4] $$S = \sqrt{\frac{1}{mN-K}\sum_{j=1}^{N}\sum_{p\in B,T}\sum_{i\in G_j^p}\left(Y_{ij} - \bar{Y}_{.j}^{p}\right)^2}\ ,$$

where *N* is the total number of timepoints, *K* is a degrees-of-freedom correction, and G_j^p indicates which cases are in condition *p* at time point *j*, for *j* = 1,…, *N* and *p* = B for Baseline, T for Treatment. Finally, HomVEE applies the small sample correction and estimates the standard error of the small-sample corrected D-CES following equations 7 and 8, respectively, in Shadish, Hedges, and Pustjeovsky (2014).

References

Berk, R.A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83,* 460– 472.

Hartmann, D.P., Barrios, B.A., & Wood, D.D. (2004). Principles of behavioral observation. In S. N. Haynes & E. M. Hieby (Eds.), *Comprehensive handbook of psychological assessment. Vol. 3: Behavioral assessment* (pp. 108–127). New York, NY: Wiley.

Hedges, L.V., Pustejovsky, J.E., & Shadish, W.A. (2012). A standardized mean difference effect size for single case designs. *Journal of Research Synthesis Methods, 3*, 224–239.

Hedges, L.V., Pustejovsky, J.E., & Shadish, W.A. (2013). A standardized mean difference effect size for multiple baseline designs. *Journal of Research Synthesis Methods, 4,* 324–341.

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Expanding analysis and use of single-case research. *Education and Treatment of Children, 35,* 269–290.

Pustejovsky, J.E., Hedges, L.V., & Shadish, W.L. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39,* 368–393.

Shadish, W.R., Hedges, L.V., & Pustejovsky, J.E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147.

Suen, H.K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

What Works Clearinghouse. (2020). Handbooks and Other Resources: Procedures and Standards Handbooks. Retrieved June 4, 2020, from https://ies.ed.gov/ncee/wwc/handbooks

Appendix E

Handling of missing data and imputation, from the WWC 4.1 Standards Handbook

This page has been left blank for double sided copying.

This appendix replicates the What Works Clearinghouse Version 4.1 standards for research with this design, except for minor wording changes to tailor them to the HomVEE context. [84]

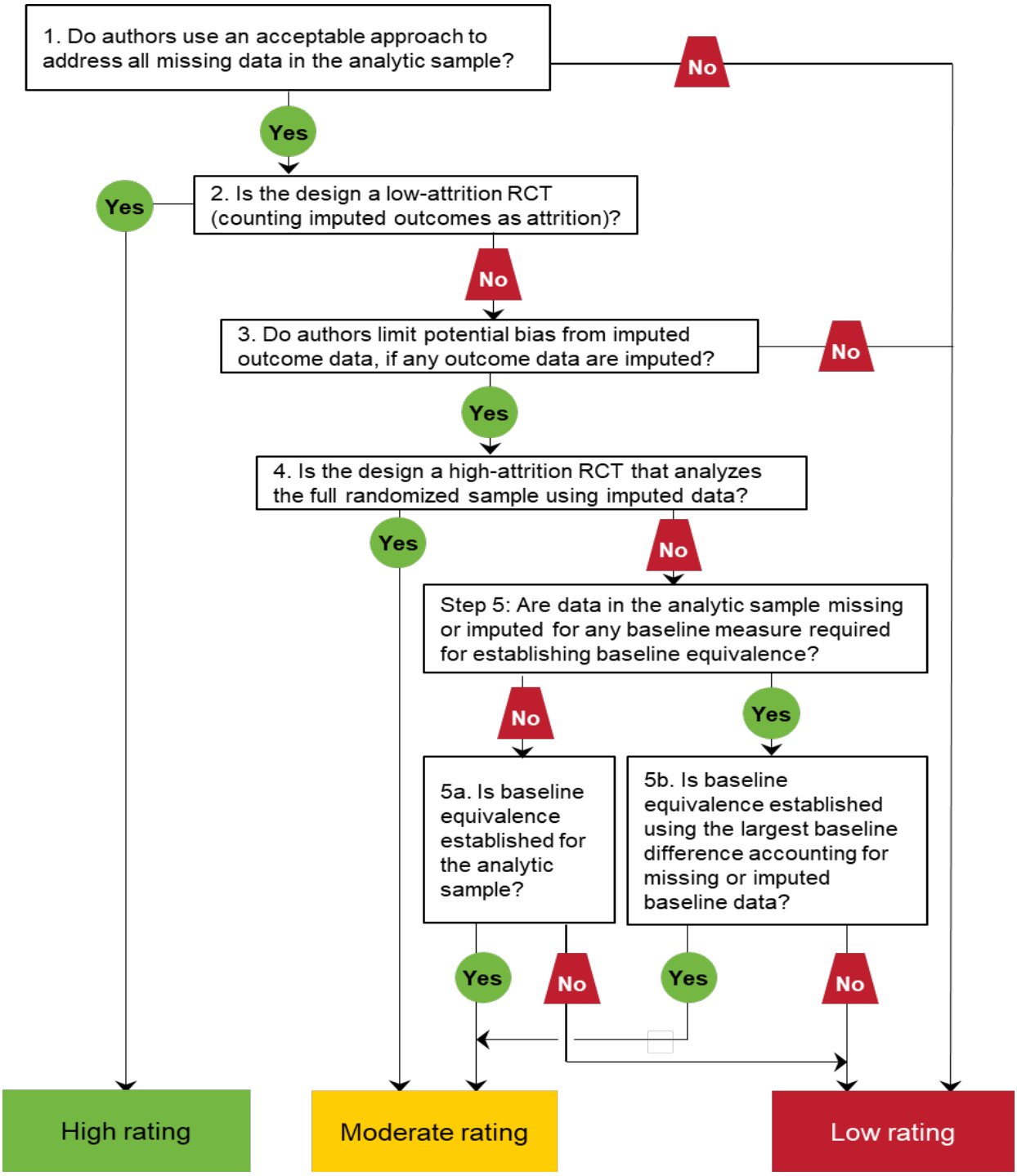
A. Analyses with missing data

Despite the best efforts of researchers, sometimes it is not possible to collect data for all participants in a study sample. Authors might use a variety of analytical approaches to address missing data for baseline or outcome measures. For example, a manuscript might focus on the analytic sample of participants for which all data were collected, or the authors may impute values for the missing data so that more participants can be included in the analysis. The review process for a manuscript about a study with missing data depends on the study design (randomized controlled trial [RCT] or non-experimental comparison group design [NED]), the method used to address the missing data, and whether the sample examined in the manuscript has missing baseline data, outcome data, or both. This review process applies to RCTs and NEDs. It does not apply to regression discontinuity designs because the review process for those designs has requirements on missing data and sample loss that are specific to those designs. Also, this review process does not apply to single case design because these designs do not experience sample loss.

The steps in the review process for RCTs and NEDs with missing data are outlined in Figure E.1. Steps 1 and 2 must be performed for any manuscript about a study with missing data, Steps 3 and 4 relate to manuscripts with imputed outcome data in the analytic sample, and Step 5 relates to manuscripts with imputed or missing baseline data in the analytic sample. We describe each of these steps in detail next.

[84] What Works Clearinghouse. (2020). Handbooks and Other Resources: Procedures and Standards Handbooks. Retrieved June 4, 2020, from https://ies.ed.gov/ncee/wwc/handbooks.

Figure E.1. Manuscript ratings for randomized controlled trials and non-experimental comparison group designs with missing outcome or baseline data



Note: To receive a rating of high or moderate, the manuscript must also satisfy the requirements in Chapter III, including, but not limited to, that the manuscript must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Step 1. Do authors use an acceptable approach to address all missing data in the analytic sample?

The first step in the review process for manuscripts with missing data is to determine whether any imputed data used in the analysis were generated using an acceptable imputation method. To be eligible to be rated moderate or high, an analysis must use one of the methods described in Table E.1 to address the missing data. This requirement applies to all data used in the analysis, whether for an outcome measure or a baseline measure. More specifically, the requirement applies both to baseline measures that are required for assessing baseline equivalence and those that are not.

Analyses that include any imputed outcome or baseline data based on other approaches not listed in Exhibit III.19 are rated low.

When an analysis uses one or more of these methods and satisfies all other requirements to receive a rating of moderate or high, HomVEE will report findings, including effect sizes, according to the general approach to HomVEE reporting outlined in Chapter II of this handbook. However, HomVEE will not report statistical significance for methods that do not provide accurate standard error estimates. For some other methods, HomVEE will report statistical significance provided certain requirements are met, as described in the last column in Table E.1.

All but one of the acceptable approaches in Table E.1 can provide unbiased estimates of the effectiveness of an intervention based on the assumption that the missing data do not depend on unmeasured factors. The exception is complete case analysis, which requires a more restrictive assumption that the missing data also do not depend on measured factors. Because of this, many researchers have recommended against using complete case analysis to address missing data (for example, Little et al., 2012; Peugh & Enders, 2004). Nevertheless, HomVEE considers complete case analysis to be an acceptable approach for addressing missing data because possible bias due to measured factors can be assessed through the attrition standard and HomVEE's baseline equivalence requirement, as described in Chapter III, Sections B.1 and B.2 of this handbook, respectively.

In addition, Jones (1996) and Allison (2002) raised concerns about using the approach in the last row of Table E.1, imputation to a constant combined with including a missing data indicator, outside of RCTs. Consequently, HomVEE considers this approach acceptable for any baseline data in RCTs regardless of their sample attrition. However, in a NED or compromised RCT, the approach is acceptable only when applied to baseline measures that HomVEE does not require for assessing baseline equivalence.

To obtain appropriate estimates of statistical significance in manuscripts about cluster-level assignment studies that analyze individual-level data, approaches to address missing outcome data must account for the correlation of outcomes within clusters. This can be done using standard approaches in complete case analyses. However, as noted in the last column of Table E.1, for HomVEE to confirm statistical significance in a manuscript about a study with cluster-level assignment that uses regression imputation, maximum likelihood, or nonresponse weights to address missing outcome data, and analyzes individual-level data, the manuscript must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness. In analyses using these three approaches that do not include an acceptable adjustment, HomVEE will not apply its adjustment for clustering, as described in Chapter III of this handbook, because it may not be accurate for analyses using these methods. HomVEE does not currently have a recommended method of calculating standard errors in these analyses of cluster-level assignment studies, and the burden for demonstrating that the approach is appropriate rests with the authors.

The training for HomVEE reviewers provides an overview of the approaches listed in Table E.1 and on the expectation that these approaches must be applied so that the findings from analyses that include imputed data meet HomVEE standards. Reviewers are instructed to bring questions on whether a manuscript about a study appropriately applied any of these methods to the review team leadership.

Finally, if a manuscript about study uses an approach not listed in Table E.1, HomVEE may consider it an acceptable approach after consultation with experts if the following is also true:

- The approach is supported with a citation to a peer-reviewed journal article or textbook that describes the procedure.
- The cited research demonstrates that the approach can produce unbiased estimates under an assumption that the missing data are unrelated to unmeasured factors.,

Table E.1. Acceptable approaches for addressing missing baseline or outcome data

Approach	Description	Requirements	Statistical significance
Complete case analysis.	Exclusion of observations with missing outcome and/or baseline data from the analysis.	None.	HomVEE has no additional requirements for reporting statistical significance from analyses that use this method.
Regression imputation.	A regression model to predict imputed values for the missing data. This includes estimating imputed values from a single regression model, and multiple imputation, which involves generating multiple datasets that contain imputed values for missing data through the repeated application of an imputation algorithm, such as chained equations.	The imputation regression model must: **1.** Be conducted separately for the intervention and comparison groups or include an indicator variable for intervention status, **2.** Include all of the covariates that are used for statistical adjustment in the impact estimation model, and **3.** Include the outcome when imputing missing baseline data.	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or multiple imputation. For multiple imputation, the statistical significance calculation must: **1.** Be based on at least five sets of imputations, and **2.** Account for (1) the within-imputation variance component, (2) the between-imputation variance component, and (3) the number of imputations. Most established multiple imputation routines satisfy this requirement. Additionally, a manuscript about a cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Maximum likelihood.	An iterative routine to estimate model parameters and impute values for the missing data. Some examples are the expectation-maximization algorithm and full information maximum likelihood.	The procedure must use a standard statistical package or be supported with a citation to a peer-reviewed methodological journal article or textbook.	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or estimates based on the information matrix. Additionally, a manuscript about a cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.

Approach	Description	Requirements	Statistical significance
Nonresponse weights.	Use of weights based on estimated probabilities of having a nonmissing outcome, yielding greater weight for participants with a higher probability of having missing outcome data. For example, the probabilities may be estimated from a logit or probit model.	Acceptable only for missing outcome data, not for missing baseline data. The estimated probabilities used to construct the weights must: **1.** Be estimated separately for the intervention and comparison groups or include an indicator variable for intervention status, and **2.** Include all baseline measures that are required for baseline equivalence. Including additional covariates is acceptable but not required because doing so may lead to less precise impact estimates without providing a substantial reduction in bias.	The analysis must properly account for the stratified sampling associated with the weights (as discussed in Wooldridge (2002), p. 594). Additionally, a manuscript about a cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Replacing missing data with a constant combined with including a missing data indicator.	Setting all missing values for a baseline measure to a single value, and including an indicator variable for records missing data on the measure in the impact estimation model.	Acceptable only for missing baseline data, not for missing outcome data. When applied to a baseline measure required for assessing baseline equivalence, the method is acceptable only in RCTs regardless of sample attrition, but not in NEDs or compromised RCTs.	HomVEE has no additional requirements for reporting statistical significance from analyses that use this method.

Note: Requirements in this table are based on recommendations in several sources, including Allison (2002), Azur, Stuart, Frangakis, and Leaf (2011); Little and Rubin (2002); Puma, Olsen, Bell, and Price (2009); Rubin (1987); Schafer (1999); and Wooldridge (2002).

HomVEE review process for Step 1 of the review of manuscripts about studies with missing data

- If the manuscript uses an acceptable approach to address all missing data in the analytic sample, then continue to Step 2.
- If the manuscript does not use an acceptable approach to address all missing data in the analytic sample, then the manuscript is rated low.

Step 2. Is the design a low-attrition randomized controlled trial (counting imputed outcomes as attrition)?

The second step in the review process for manuscripts with missing data is to determine whether the manuscript reports a low-attrition RCT as described in Chapter III, Section B.1. When calculating overall and differential attrition rates, sample members with imputed outcome data are counted as missing because both missing and imputed data represent a potential threat of bias. The use of imputed data can mitigate that bias if the missing data do not depend on unmeasured factors, but otherwise may not. When attrition is low, HomVEE will ignore the potential bias from imputed data because the amount of missing or imputed data is unlikely to lead to bias that exceeds HomVEE's tolerable level of potential bias. A low-attrition RCT is eligible to be rated high as long as the authors used an acceptable method to address missing data.

WWC review process for Step 2 of the review of manuscripts about studies with missing data

- If the study is a low-attrition RCT, then the study is eligible to receive the rating high. To receive this rating, the manuscript must also satisfy the requirements in Chapter III, including that the it must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the study is a NED, high-attrition RCT, or compromised RCT, then continue to Step 3 of the review process for manuscripts about studies with missing data.

Step 3. Do authors limit potential bias from imputed outcome data, if any outcome data are imputed?

Imputed outcome data can affect the rating of a NED, high-attrition RCT, or compromised RCT in two ways. The first of these is addressed in this step. To be eligible for a rating of moderate, NEDs, high-attrition RCTs, and compromised RCTs with imputed outcome data in the analytic sample must satisfy an additional requirement designed to limit potential bias from using imputed outcome data instead of actual outcome data.

The imputation methods HomVEE considers acceptable are based on an assumption that the missing data depend on measured factors, not unmeasured factors. If that assumption does not hold, then impact estimates may be biased. Therefore, manuscripts about group design studies besides low-attrition RCTs that use acceptable approaches to impute outcome data must demonstrate that they limit the potential bias from using imputed data to measure impacts to less than 0.05 standard deviation as described in this step.

An analysis of a sample with imputed outcome data can produce biased estimates of the effect of the intervention if the participants with observed data differ from the participants with missing data, and some of the differences are unmeasured. In this case, if outcomes could be obtained for all sample members, then the average for participants in the intervention or comparison condition with observed outcome data would differ from the average for participants whose outcome data were not observed. Comparing the differences in these means for the intervention and comparison groups, if known, would indicate the magnitude of possible bias, but because the missing outcomes are not observed, HomVEE instead assesses the bias using baseline data.

HomVEE estimates the potential bias from missing outcome data due to unmeasured factors by comparing means of the baseline measure required for assessing baseline equivalence, separately for the intervention and comparison groups, for two samples: the complete analytic sample and the analytic sample restricted to cases with observed outcome data. A smaller difference in these two means within one or both conditions lowers the likelihood that the missing data are related to factors that could lead to bias in the impact estimate.

To translate the intervention and comparison group differences in baseline means into an estimate of bias in the outcome effect size, HomVEE uses the pooled standard deviation of the baseline measure and the correlation between the baseline and outcome measure. Section B of this appendix (Appendix E) provides the formulas HomVEE uses to estimate the potential bias (equations E.5.0–E.5.2). Section B of this appendix also describes the approach used when baseline equivalence must be assessed on multiple baseline measures, as is the case in HomVEE. The formulas used to assess the bias also differ depending on whether the baseline measure is observed for all participants in the analytic sample (equations E.10.0–E.10.2 in this appendix).

- **When the baseline measure is observed for all participants in the analytic sample**, HomVEE requires the following data from the authors: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups—these are the same data used to assess baseline equivalence; (b) the means of the baseline measure for the participants in the analytic sample with observed outcome data, separately for the intervention and comparison groups; and (c) the correlation between the baseline and the outcome measures. The correlation can be estimated on a sample other than the analytic sample, such as the complete case sample, or from data from outside the study if a subject matter expert judges the settings to be similar. However, the correlation must not be estimated using imputed data.
- **When the baseline measure is imputed or missing for some participants in the analytic sample**, in addition to (c), the following data are required: (d) the means of the baseline measure for the participants in the analytic sample with observed baseline data, separately for the intervention and comparison groups; (e) the means of the baseline measure for the participants in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (f) the standard deviations of the baseline measure for either the sample of participants in the analytic sample with observed baseline data or the sample with observed baseline and outcome data; and (g) the number of participants with observed baseline data in the analytic sample by condition.

If these data are not reported in the manuscript, then HomVEE will request them from the authors. There are two special considerations for applying the requirement in Step 3 when an analysis uses nonresponse weights or complete case analysis:

- An analysis that uses nonresponse weights to address missing outcome data must also satisfy the requirement to limit the potential bias from using imputed data. For these analyses, separately for the intervention and comparison groups, HomVEE compares a different pair of means of the baseline measure. Instead of the complete analytic sample, which for a nonresponse weighted analysis would be restricted to cases with observed outcome data, HomVEE uses the sample used to estimate the weights, including cases with missing outcome data. The second mean remains the sample with observed outcome data.
- A complete case analysis that addresses missing data by excluding cases with missing outcome data, rather than imputing it, does not need to satisfy this requirement. The exclusion of complete case analyses from this requirement is not intended to imply that complete case analyses are believed to be a stronger approach for addressing missing data. Rather, HomVEE's approach recognizes that the attrition standard and baseline equivalence requirement can limit bias in complete case analyses because the missing data affect the analytic sample.

HomVEE review process for Step 3 of the review of manuscripts about studies with missing data

- If the authors limit potential bias from imputed outcome data, as assessed using the formulas in this Appendix E, or the analytic sample contains no imputed outcome data, then continue to Step 4 of the review process for manuscripts about studies with missing data.
- If the authors do not limit potential bias from unmeasured factors, then the manuscript is rated low.

Step 4. Is the design a high-attrition RCT that analyzes the full randomized sample using imputed data?

The fourth step in the review process for missing outcome data addresses a second way imputed outcome data can affect the rating of a manuscript. When authors analyze a high- attrition RCT by imputing outcome data so that they analyze the full sample that was randomized to conditions, the manuscript does not need to satisfy the baseline equivalence requirement to be eligible to receive the rating moderate.

In general, HomVEE requires that high-attrition RCTs satisfy the baseline equivalence requirement because of a risk of bias from compositional differences between the remaining intervention and comparison group members. However, some high-attrition RCTs impute all missing outcome data and analyze the original randomized sample. These high-attrition RCTs do not need to satisfy the baseline equivalence requirement because of a presumption that intervention and comparison groups that result from random assignment are unlikely to have substantive compositional differences. Imputing missing outcome data and analyzing the full randomized sample preserves the integrity of the originally randomized groups. Although compositional differences are not considered a threat to bias, like other high-attrition RCTs, manuscripts about these studies are eligible to be rated only moderate. These manuscripts are not eligible for the highest rating because of the risk of bias from imputing a larger amount of missing outcome data compared with a low-attrition RCT.

All NEDs, high-attrition RCTs that do not analyze the original randomized sample, and compromised RCTs must satisfy the baseline equivalence requirement (Step 5 in Figure E.1).

HomVEE review process for Step 4 of the review of manuscripts about studies with missing data

- If the design is a QED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, and the analytic sample does not include missing or imputed data for any baseline measure required for establishing baseline equivalence, then continue to Step 5a of the review process for studies with missing data.
- If the design is a QED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, and the analytic sample includes some missing or imputed data for a baseline measure required for establishing baseline equivalence, then continue to Step 5b of the review process for studies with missing data.

Step 5. Are data in the analytic sample missing or imputed for any baseline measure required for establishing baseline equivalence?

NEDs, high-attrition RCTs that do not impute data to analyze the full randomized sample, and compromised RCTs must satisfy the baseline equivalence requirement to be eligible to be rated moderate. However, it is not possible for HomVEE to assess baseline equivalence on the full analytic sample using actual data when some data are missing or imputed for a measure required for assessing baseline equivalence.

HomVEE review process for Step 5 of the review of manuscripts about studies with missing data

- If the manuscript is about a high-attrition RCT that analyzes the original randomized sample, then the manuscript is eligible to receive the rating moderate and does not need to satisfy the baseline equivalence requirement. To receive this rating, the manuscript must also satisfy the requirements in Chapter III, including that the manuscript must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the manuscript is about a study that is a NED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, then the manuscript must satisfy the baseline equivalence requirement to be eligible to receive the rating moderate. Continue to Step 5 of the review process for manuscripts with missing data.

Step 5a. Is baseline equivalence established for the analytic sample?

If all of the missing or imputed baseline data in the analytic sample are for baseline measures not required for satisfying baseline equivalence, or no baseline data are missing or imputed, then baseline equivalence can be assessed using the usual approach described in Chapter III, Section B.2. A manuscript that satisfies the baseline equivalence requirement using actual data for the analytic sample is eligible to be rated moderate.

HomVEE review process for Step 5a of the review of manuscripts about studies with missing data

- If the manuscript satisfies the baseline equivalence requirement using actual baseline data, the manuscript is eligible to receive the rating moderate. To receive this rating, the manuscript must also satisfy the requirements in Chapter III, including that the manuscript must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the manuscript does not satisfy the baseline equivalence requirement using actual baseline data, the manuscript is rated low.

An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.

Step 5b. Is baseline equivalence established using the largest baseline difference accounting for missing or imputed baseline data?

If some data are missing or imputed for a baseline measure that is required for satisfying baseline equivalence, then HomVEE uses a different process to assess baseline equivalence. In this case, HomVEE estimates how large the baseline difference might be under different assumptions about how the missing data are related to measured or unmeasured factors. The largest of these estimates in absolute value is used as the baseline difference for the manuscript.

Just as for manuscripts about studies with complete baseline data, a manuscript about a study with missing or imputed data for a required baseline measure is eligible to be rated *moderate* if the largest estimated standardized baseline difference does not exceed 0.25 standard deviation when the analysis includes an acceptable adjustment for the baseline measure, or 0.05 standard deviation otherwise. A manuscript that satisfies this alternative baseline equivalence requirement is eligible to be rated *moderate*.

HomVEE's approach to estimating the baseline difference in manuscripts about studies with missing or imputed baseline data is similar to the approach used to estimate bias from using imputed outcome data, described above. Instead of comparing means of the baseline measure, HomVEE compares means of the outcome measure, separately for the intervention and comparison groups, for two samples: the analytic sample and the analytic sample restricted to cases with observed baseline data. A larger absolute difference in these means within a group indicates that the data may be missing in a way that is related to unmeasured sample characteristics, and the measured impact of the intervention may be biased.

To translate the intervention and comparison group differences in outcome means into an estimate of a baseline effect size, HomVEE uses the pooled standard deviation of the outcome measure and the correlation between the baseline and outcome measure. Section B of this appendix provides the formulas HomVEE uses to estimate the baseline effect size (equations E.15.0–E.15.3, E.17.0– E.17.3, E.21.0–E21.3, and E.23.0–E.23.3). If baseline equivalence must be assessed on multiple baseline measures, the formulas must be applied to each required baseline measure. The formulas used to estimate the baseline difference vary based on two factors: whether the outcome measure is observed for all participants in the analytic sample and whether the outcome data are missing or imputed.

- **When the outcome measure is observed for all participants in the analytic sample**, HomVEE requires the following data from the authors: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups; (b) the means

of the outcome measure for the participants in the analytic sample with observed baseline data, separately for the intervention and comparison groups; (c) the correlation between the baseline and the outcome measures; and (d) an estimate of the baseline difference based on study data. As noted in Step 3 of the section on imputed outcome data, the correlation can be estimated on a sample other than the analytic sample but must not be estimated using imputed data. If the authors did not impute the baseline data, then HomVEE will use baseline means and standard deviations to measure the baseline difference for the portion of the analytic sample with observed baseline data. However, if the authors did impute baseline data, then HomVEE will include the imputed data when calculating the means but will use standard deviations based only on the observed data.

- **When the outcome measure is imputed for some participants in the analytic sample**, in addition to (c) and (d) listed in the previous bullet point, the following data are required: (e) the means of the outcome measure for the participants in the analytic sample with observed outcome data, separately for the intervention and comparison groups; (f) the means of the outcome measure for the participants in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (g) the standard deviations of the outcome measure for either the sample of participants in the analytic sample with observed outcome data or the sample with observed baseline and outcome data; and (h) the number of participants with observed outcome data in the analytic sample by condition.

If these data are not reported in the manuscript, then HomVEE will request them from the authors.

The two special considerations for applying the requirement in Step 5b when an analysis uses nonresponse weights or complete case analysis are as follows:

- An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.
- Because no baseline data are missing or imputed, a complete case analysis that excludes cases with missing baseline data must satisfy the baseline equivalence requirement using the observed data for the analytic sample, as described in Chapter III, Section B.2, rather than using the formulas in Section B of this Appendix E. In other words, the complete case analysis must satisfy baseline equivalence using the Step 5a described here and not Step 5b.

HomVEE review process for Step 5b of the review of manuscripts about studies with missing data

- If the baseline equivalence requirement is satisfied using the largest baseline difference (estimated according to the formulas in Section 2 of this Appendix E) accounting for the missing or imputed data, the study is eligible to receive a rating of moderate. To receive this rating, the design must also satisfy the requirements in Chapter III, including, but not limited to, that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the baseline equivalence requirement is not satisfied using the largest baseline difference accounting for the missing or imputed data, then the study is rated low.

B. Assessing the bias when the baseline measure is observed for all participants in the analytic sample

The imputation methods that the WWC and HomVEE consider acceptable require assuming that data are missing at random (MAR), which means the missing data depend on measured factors but not on unmeasured factors. If that assumption does not hold, then the impact estimates may be biased. Therefore, non-experimental group designs (NEDs) and high-attrition randomized controlled trials (RCTs) that use acceptable approaches to impute outcome data must demonstrate that they limit the potential bias from using imputed data to measure impacts. Specifically, potential bias due to deviations from the MAR assumption must not exceed 0.05 standard deviation.

HomVEE uses a proxy pattern-mixture modeling approach to estimate the largest possible bias in an impact estimate under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

To bound the bias, we begin by specifying that the probability that we observe an outcome for a given subject is related to the baseline measure and the outcome, which is unmeasured for some cases. This probability in the intervention group $\left(j = i\right)$ or comparison group $\left(j = c\right)$ is given by the following function *m*:

[E.1] $$P_j\left(x, y\right) = m\left(\frac{x}{s_x} + \lambda_j \frac{y}{s_y}\right),$$

where *x* is the baseline measure for a subject, *y* is the outcome measure for the subject, s_x and s_y are the standard deviations of the baseline and outcome measures, and λ_j measures the deviations from the MAR assumption for group *j*. When $\lambda_j = 0$, the MAR assumption holds for group *j* because the missing data depend only on measured baseline data. As λ_j increases, the missingness depends more strongly on the outcome, which may be unmeasured.

Following Andridge and Little (2011), we can write the unmeasured full-sample outcome mean in a group $\left(\overline{y}_j\right)$ as a function of the complete case outcome mean $\left(\overline{y}_{jR}\right)$, the full-sample and complete case baseline means ($\overline{x}_j$ and $\overline{x}_{jR}$), and the correlation between the outcome and the baseline measure ρ:

[E.2.0] $$\overline{y}_j = \overline{y}_{jR} + f_j\left(\rho\right)\frac{s_y}{s_x}\left[\overline{x}_j - \overline{x}_{jR}\right],$$

where the function of ρ is assumed to be:

[E.2.1] $$f_j\left(\rho\right) = \frac{\lambda_j + \rho}{\lambda_j \rho + 1}.$$

In many cases, the value of $\overline{y}_j$ will deviate more from the observed mean of $\overline{y}_{jR}$ when there is a larger absolute difference between the full-sample and complete case baseline means. Intuitively, this is because a larger difference means that the subjects with missing outcome data appear different from those with observed outcomes.

When MAR holds, $f_i(\rho) = f_c(\rho) = \rho$ (because $\lambda_i = \lambda_c = 0$), and the expected value of $\overline{y}_j$ is equal to what a researcher would obtain for the full-sample outcome mean when imputing missing values of the outcome measure with predicted values from a regression of the outcome on the baseline measure. But as λ_i or λ_c become larger, the value of $f_j(\rho)$ becomes larger (approaching $1/\rho$), and the outcome mean for the full sample will deviate from the researcher's estimate of the mean using imputed data.

The effect size obtained using an imputation method based on the MAR assumption can be written as the difference in the estimated full-sample intervention and comparison group outcome means with an adjustment for the baseline measure, given by:

[E.3.0] $$g_{MAR} = \frac{1}{s_y}\left(\left\{y_{iR} + c\left[\overline{x}_i - \overline{x}_{iR}\right]\right\} - \left\{y_{cR} + c\left[\overline{x}_c - \overline{x}_{cR}\right]\right\} - c\left[\overline{x}_i - \overline{x}_c\right]\right),$$

where *c* is the coefficient from a regression of *y* on *x*, and is equal to $\rho\left(s_y / s_x\right)$.

But this equation can be generalized to the case where the MAR assumption does not hold:

[E.3.1] $$g_{NMAR} = \frac{1}{s_y}\left(\left\{y_{iR} + f_i(\rho)\frac{s_y}{s_x}\left[\overline{x}_i - \overline{x}_{iR}\right]\right\} - \left\{y_{cR} + f_c(\rho)\frac{s_y}{s_x}\left[\overline{x}_c - \overline{x}_{cR}\right]\right\} - c\left[\overline{x}_i - \overline{x}_c\right]\right).$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the MAR assumption:

[E.4] $$Bias_y = \frac{1}{s_x}\left\{\left(f_i(\rho) - \rho\right)\left[\overline{x}_i - \overline{x}_{iR}\right] - \left(f_c(\rho) - \rho\right)\left[\overline{x}_c - \overline{x}_{cR}\right]\right\}.$$

Because $f_j(\rho)$ is bounded between ρ and $1/\rho$, the largest bias, in absolute value, due to deviations from the MAR assumption is given by the maximum of the values given by the following three equations:

[E.5.0] $$B1 = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left[\overline{x}_c - \overline{x}_{cR}\right]\right|$$

[E.5.1] $$B2 = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left[\overline{x}_i - \overline{x}_{iR}\right]\right|$$

[E.5.2] $$B3 = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left[\left(\overline{x}_i - \overline{x}_{iR}\right) - \left(\overline{x}_c - \overline{x}_{cR}\right)\right]\right|.$$

The bounds in equations E.5.0, E.5.1, and E.5.2 will be calculated using data reported in manuscripts or obtained from authors. The equations include the following data elements: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups ($\overline{x}_i$, $\overline{x}_c$, and the standard deviations are used to calculate the pooled within-group standard deviation s_x [85]); (b) the means of the baseline measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups ($\overline{x}_{iR}$, $\overline{x}_{cR}$); and (c) the correlation between the baseline and the outcome measures (ρ). We have applied a simple correction for bias in the unadjusted Hedges' *g* effect size when the sample size is small, developed by Hedges (1981), which produces an unbiased effect size estimate by multiplying Hedges' *g* by a factor of $\omega = \left[1 - 3/\left(4N - 9\right)\right]$, with *N* being the total sample size.

For simplicity, these bounds were derived for a single baseline measure. If multiple baseline measures were used to form the imputed values in a manuscript, it is acceptable, but not required, to replace the baseline means with the average predicted value of the outcome, that is, the average of the values used to make adjustments to the outcome measure to produce an adjusted mean. In this case, $1/s_x$ is removed from the calculation of the bounds and replaced with $1/s_y$ because the predicted values have units of the dependent variable. Additionally, when baseline equivalence is required on multiple baseline measures, the imputed values must adjust for all baseline measures (as required for establishing baseline equivalence) and that the bounds are calculated using the average of the predicted values.

C. Assessing the bias when the baseline measure is imputed or missing for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds in equations E.5.0–E.5.2. This is because the means of the baseline measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed outcome data.

Instead, the bounds can be calculated using equations E.10.0–E.10.2. These bounds can be derived by first writing the full sample outcome mean as a weighted sum of the outcome mean for the sample with missing data on the baseline measure, and the sample with observed data on the baseline measure:

[E.6.0] $$\overline{y}_j = \left(\frac{n_j - n_{jx}}{n_j}\right)\overline{y}_{j\sim x} + \left(\frac{n_{jx}}{n_j}\right)\overline{y}_{jx},$$

where n_j is the number of observations in the analytic sample for group *j*, n_{jx} is the number of observations in the analytic sample for group *j* with an observed value of the baseline measure, $\overline{y}_{j\sim x}$ is the

[85] HomVEE will use the same procedures that the WWC uses to calculate pooled standard deviation, as indicated in section IV.A of the WWC Procedures Handbook, Version 4.1 (U.S. Department of Education 2020a) .

outcome mean for the observations in the analytic sample for group *j* missing the baseline measure, and $\overline{y}_{jx}$ is the outcome mean for the remaining members of the analytic sample for group *j*.

We assume that the analytic sample includes no cases where both the baseline and outcome data are missing, so $\overline{y}_{j\sim x}$ is observed. But $\overline{y}_{jx}$ is not observed because some cases with observed baseline data have missing outcome data. To address this, we write $\overline{y}_{jx}$ as a function of observed measures:

$$[E.6.1] \qquad \overline{y}_j = \left(\frac{n_j - n_{jx}}{n_j}\right)\overline{y}_{j\sim x} + \left(\frac{n_{jx}}{n_j}\right)\left(\overline{y}_{jxy} + f_j(\rho)\frac{s_y}{s_x}\left[\overline{x}_{jx} - \overline{x}_{jxy}\right]\right),$$

where $\overline{y}_{jxy}$ is the outcome mean for the observations in the complete case analytic sample for group *j* observed at both baseline and for the collection of outcomes, $\overline{x}_{jxy}$ is the baseline mean for the same sample, and $\overline{x}_{jx}$ is the baseline mean for the sample with observed baseline data but possibly missing outcome data. This equation can be rewritten as:

$$[E.6.2] \qquad \overline{y}_j = \overline{y}_{jxy} + \left(\frac{n_j - n_{jx}}{n_j}\right)\left[\overline{y}_{j\sim x} - \overline{y}_{jxy}\right] + \left(\frac{n_{jx}}{n_j}\right) f_j(\rho)\frac{s_y}{s_x}\left[\overline{x}_{jx} - \overline{x}_{jxy}\right].$$

The effect size obtained using an imputation method based on the MAR assumption $\left(f_j(\rho) = \rho\right)$ can be written as the difference in the estimated full-sample intervention and comparison group outcome means,[86] given by:

$$[E.7] \qquad g_{MAR} = \frac{1}{s_y}\left(\left\{\overline{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i}\right)\left[\overline{y}_{i\sim x} - \overline{y}_{ixy}\right] + \left(\frac{n_{ix}}{n_i}\right)c\left[\overline{x}_{ix} - \overline{x}_{ixy}\right]\right\} - \left\{\overline{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c}\right)\left[\overline{y}_{c\sim x} - \overline{y}_{cxy}\right] + \left(\frac{n_{cx}}{n_c}\right)c\left[\overline{x}_{cx} - \overline{x}_{cxy}\right]\right\}\right).$$

[86] In this equation, we ignore an adjustment for the baseline measure. Because the baseline data are imputed, deviations from the MAR assumption can lead to bias in this adjustment. This source of potential bias in the outcome effect size is accounted for separately through the baseline equivalence requirement when data are missing.

The more general equation that allows deviations from the MAR assumption is given by:

[E.8]
$$g_{NMAR} = \frac{1}{s_y}\left(\left\{\bar{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i}\right)\left[\bar{y}_{i\sim x} - \bar{y}_{ixy}\right] + \left(\frac{n_{ix}}{n_i}\right) f_i(\rho)\frac{s_y}{s_x}\left[\bar{x}_{ix} - \bar{x}_{ixy}\right]\right\} - \left\{\bar{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c}\right)\left[\bar{y}_{c\sim x} - \bar{y}_{cxy}\right] + \left(\frac{n_{cx}}{n_c}\right) f_c(\rho)\frac{s_y}{s_x}\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right\}\right).$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the MAR assumption:

[E.9]
$$Bias_y = \frac{1}{s_x}\left\{\left(\frac{n_{ix}}{n_i}\right)\left(f_i(\rho) - \rho\right)\left[\bar{x}_{ix} - \bar{x}_{ixy}\right] - \left(\frac{n_{cx}}{n_c}\right)\left(f_c(\rho) - \rho\right)\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right\}.$$

The absolute value of this bias is no greater than the maximum of $B1^* - B3^*$:

[E.10.0]
$$B1^* = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left(\frac{n_{ix}}{n_i}\right)\left[\bar{x}_{ix} - \bar{x}_{ixy}\right]\right|$$

[E.10.1]
$$B2^* = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left(\frac{n_{cx}}{n_c}\right)\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right|$$

[E.10.2]
$$B3^* = \omega\left|\frac{1}{s_x}\frac{1-\rho^2}{\rho}\left[\left(\frac{n_{ix}}{n_i}\right)\left(\bar{x}_{ix} - \bar{x}_{ixy}\right) - \left(\frac{n_{cx}}{n_c}\right)\left(\bar{x}_{cx} - \bar{x}_{cxy}\right)\right]\right|.$$

In addition to the correlation between the baseline and the outcome measures (ρ) used in calculating $B1 - B3$ discussed above, the bounds in equations E.10.0–E.10.2 include the following data elements: (1) the means of the baseline measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups ($\bar{x}_{ix}$, and $\bar{x}_{cx}$); (2) the means of the baseline measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups ($\bar{x}_{ixy}$, and $\bar{x}_{cxy}$); (3) the standard deviations of the baseline measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data, separately for the intervention and comparison groups, which are used to calculate s_x [87]; and (4) the number of subjects with observed baseline data in the analytic sample by condition (n_{cx}).

The formulas for $B1^* - B3^*$ reduce to $B1 - B3$ when there are no missing baseline data.

[87] For simplicity, this is referred to using the consistent notation despite the difference in the data used to calculate it.

D. Bounding the baseline difference when the outcome is observed for all subjects in the analytic sample

It is not possible to assess baseline equivalence using observed data for the analytic sample in non-experimental group designs (NEDs) and high-attrition randomized controlled trials (RCTs) that use acceptable approaches to impute baseline data or are missing some baseline data for the analytic sample. However, as the WWC does, HomVEE will consider the potential bias from baseline differences to be limited if, under different assumptions about whether the data are missing at random (MAR), the standardized baseline difference does not exceed 0.25 standard deviation when the analysis includes an acceptable adjustment for the baseline measure, or 0.05 standard deviation otherwise. This requirement applies only to baseline measures that are required for satisfying the baseline equivalence requirement.

HomVEE uses the same proxy pattern-mixture modeling approach used to address imputed outcome data to estimate the largest possible baseline difference under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

Using the same notation introduced earlier in this appendix, the baseline mean for a sample with missing or imputed baseline data can be modelled using:

[E.11] $$\bar{x}_j = \bar{x}_{jR} + g_j(\rho)\frac{s_x}{s_y}\left[\bar{y}_j - \bar{y}_{jR}\right],$$

where $\bar{x}_j$ and $\bar{x}_{jR}$ are the full-sample and complete case baseline means, $\bar{y}_j$ and $\bar{y}_{jR}$ are the full-sample and complete case outcome means, ρ is the correlation between the outcome and the baseline measure, and

[E.12] $$g_j(\rho) = \frac{1}{f_j(\rho)} = \frac{\lambda_j\rho + 1}{\lambda_j\rho}.$$

The full-sample baseline effect size obtained using an imputation method based on the MAR assumption ($g_c(\rho) = g_i(\rho) = \rho$ when λ_j approaches ∞) can be written as the baseline effect size for the observed sample g_{xR} with an adjustment for the difference between the full-sample and complete case outcome means in the intervention and comparison groups, given by:

[E.13] $$g_{xMAR} = g_{xR} + \frac{\rho}{s_y}\left(\left[\bar{y}_i - \bar{y}_{iR}\right] - \left[\bar{y}_c - \bar{y}_{cR}\right]\right),$$

where $g_{xR} = \frac{1}{s_x}\left(\bar{x}_{iR} - \bar{x}_{cR}\right)$. The more general equation for the baseline effect size that allows for deviations from the MAR is:

[E.14] $$g_{xNMAR} = g_{xR} + \frac{1}{s_y}\left(g_i(\rho)\left[\overline{y}_i - \overline{y}_{iR}\right] - g_c(\rho)\left[\overline{y}_c - \overline{y}_{cR}\right]\right).$$

Because $g_j(\rho)$ is bounded between ρ and $1/\rho$, the largest baseline effect size (in absolute value) accounting for deviations from the MAR assumption is given by the maximum of the values given by the following four equations:

[E.15.0] $$C1 = \omega\left|g_{xR} + \frac{\rho}{s_y}\left(\left[\overline{y}_i - \overline{y}_{iR}\right] - \left[\overline{y}_c - \overline{y}_{cR}\right]\right)\right|$$

[E.15.1] $$C2 = \omega\left|g_{xR} + \frac{1}{\rho s_y}\left(\left[\overline{y}_i - \overline{y}_{iR}\right] - \left[\overline{y}_c - \overline{y}_{cR}\right]\right)\right|$$

[E.15.2] $$C3 = \omega\left|g_{xR} + \frac{1}{s_y}\left(\rho\left[\overline{y}_i - \overline{y}_{iR}\right] - \frac{1}{\rho}\left[\overline{y}_c - \overline{y}_{cR}\right]\right)\right|$$

[E.15.3] $$C4 = \omega\left|g_{xR} + \frac{1}{s_y}\left(\frac{1}{\rho}\left[\overline{y}_i - \overline{y}_{iR}\right] - \rho\left[\overline{y}_c - \overline{y}_{cR}\right]\right)\right|.$$

The first of these, $C1$, is $\left|g_{xMAR}\right|$, the estimate of the baseline effect size when MAR holds.

The bounds in equations E.15.0–E.15.3 will be calculated using data reported in manuscripts or obtained from authors. The equations include the following data elements: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups ($\overline{y}_i$, $\overline{y}_c$, and the standard deviations are used to calculate the pooled within-group standard deviation s_y); (b) the means of the outcome measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups ($\overline{y}_{iR}$, $\overline{y}_{cR}$); (c) the correlation between the baseline and the outcome measures (ρ); and (d) an estimate of the baseline difference based on study data (g_{xR}).

Applying the bounds in equations $C1$ – $C4$ does not require knowing the baseline effect size using imputed baseline data. Rather, these bounds use the complete case baseline effect size. When the authors impute the baseline data using an acceptable approach and the manuscript reports the baseline effect size based on imputed data, g_{xI}, a different set of bounds should be used.

Comparing g_{xMAR} and g_{xNMAR}, the bias in the imputed baseline effect size due to deviations from MAR is given by:

[E.16] $$Bias_x = \frac{1}{s_y}\left\{\left(g_i(\rho)-\rho\right)\left[\bar{y}_i - \bar{y}_{iR}\right] - \left(g_c(\rho)-\rho\right)\left[\bar{y}_c - \bar{y}_{cR}\right]\right\}.$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size:

[E.17.0] $$D1 = \omega\left|g_{xI}\right|$$

[E.17.1] $$D2 = \omega\left|g_{xI} + \frac{1}{s_y}\frac{1-\rho^2}{\rho}\left[\bar{y}_i - \bar{y}_{iR}\right]\right|$$

[E.17.2] $$D3 = \omega\left|g_{xI} - \frac{1}{s_y}\frac{1-\rho^2}{\rho}\left[\bar{y}_c - \bar{y}_{cR}\right]\right|$$

[E.17.3] $$D4 = \omega\left|g_{xI} + \frac{1}{s_y}\frac{1-\rho^2}{\rho}\left[\left(\bar{y}_i - \bar{y}_{iR}\right) - \left(\bar{y}_c - \bar{y}_{cR}\right)\right]\right|.$$

For simplicity, the bounds $C1$ – $C4$ and $D1$ – $D4$ were derived based on an imputation model based only on the relationship between the outcome and the baseline measure. If the imputation model included baseline measures in addition to the outcome, then it is acceptable but not required to replace the outcome means with the average predicted value of the baseline measure. In this case the formula should scale by s_x instead of s_y.

When baseline equivalence is required on multiple baseline measures, the bounds should be calculated separately for each baseline measure, and none may exceed the tolerable thresholds of 0.25 standard deviation when the analysis includes an acceptable adjustment, or 0.05 standard deviation otherwise.

C. Bounding the baseline difference when the outcome measure is imputed for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds $C1$ – $C4$ or $D1$ – $D4$. This is because the means of the outcome measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed baseline data.

Similar to the equation for $\bar{y}_j$ (equations E.6.0 through E.6.2), the full sample baseline mean for group j can be written as:

[E.18] $$\bar{x}_j = \bar{x}_{jxy} + \left(\frac{n_j - n_{jy}}{n_j}\right)\left[\bar{x}_{j\sim y} - \bar{x}_{jxy}\right] + \left(\frac{n_{jy}}{n_j}\right)\left(g_j(\rho)\frac{s_x}{s_y}\left[\bar{y}_{jy} - \bar{y}_{jxy}\right]\right),$$

where $\overline{x}_{jxy}$ is the baseline mean for the observations in the complete case analytic sample for group *j* and is observed at both baseline and for the collection of outcomes, $\overline{y}_{jxy}$ is the outcome mean for the same sample, and $\overline{y}_{jy}$ is the outcome mean for the sample with observed outcome data but possibly missing baseline data.

The baseline effect size obtained using an imputation method based on the MAR assumption $\left(g_j(\rho)=\rho\right)$ can be written as the difference in the estimated full-sample intervention and comparison group baseline means, given by:

[E.19]
$$g_{xMAR} = g_{xR(xy)} + \frac{1}{s_x}\left(\left\{\left(\frac{n_i - n_{iy}}{n_i}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \left(\frac{n_{iy}}{n_i}\right)\frac{\rho s_x}{s_y}\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \left(\frac{n_{cy}}{n_c}\right)\frac{\rho s_x}{s_y}\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right).$$

where $g_{xR(xy)} = \frac{1}{s_x}\left(\overline{x}_{ixy} - \overline{x}_{cxy}\right)$.

The more general formula that allows for deviations from MAR is the following:

[E.20]
$$g_{xNMAR} = g_{xR(xy)} + \frac{1}{s_x}\left(\left\{\left(\frac{n_i - n_{iy}}{n_i}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \left(\frac{n_{iy}}{n_i}\right)g_j(\rho)\frac{s_x}{s_y}\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \left(\frac{n_{cy}}{n_c}\right)g_j(\rho)\frac{s_x}{s_y}\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right).$$

The largest baseline effect size (in absolute value) accounting for deviations from the MAR assumption is given by the maximum of the values from equations E.21.0–E.21.3:

[E.21.0]
$$C1^* = \omega\left|g_{xR(xy)} + \left(\left\{\left(\frac{n_i - n_{iy}}{n_i s_x}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \rho\left(\frac{n_{iy}}{n_i s_y}\right)\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c s_x}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \rho\left(\frac{n_{cy}}{n_c s_y}\right)\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right)\right|$$

[E.21.1]
$$C2^{*} = \omega\left|g_{xR(xy)} + \left(\left\{\left(\frac{n_i - n_{iy}}{n_i s_x}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \frac{1}{\rho}\left(\frac{n_{iy}}{n_i s_y}\right)\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c s_x}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \frac{1}{\rho}\left(\frac{n_{cy}}{n_c s_y}\right)\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right)\right|$$

[E.21.2]
$$C3^{*} = \omega\left|g_{xR(xy)} + \left(\left\{\left(\frac{n_i - n_{iy}}{n_i s_x}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \rho\left(\frac{n_{iy}}{n_i s_y}\right)\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c s_x}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \frac{1}{\rho}\left(\frac{n_{cy}}{n_c s_y}\right)\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right)\right|$$

[E.21.3]
$$C4^{*} = \omega\left|g_{xR(xy)} + \left(\left\{\left(\frac{n_i - n_{iy}}{n_i s_x}\right)\left[\overline{x}_{i\sim y} - \overline{x}_{ixy}\right] + \frac{1}{\rho}\left(\frac{n_{iy}}{n_i s_y}\right)\left[\overline{y}_{iy} - \overline{y}_{ixy}\right]\right\} - \left\{\left(\frac{n_c - n_{cy}}{n_c s_x}\right)\left[\overline{x}_{c\sim y} - \overline{x}_{cxy}\right] + \rho\left(\frac{n_{cy}}{n_c s_y}\right)\left[\overline{y}_{cy} - \overline{y}_{cxy}\right]\right\}\right)\right|.$$

In addition to the correlation between the baseline and the outcome measures (ρ) and an estimate of the baseline difference based on study data (g_{xR}) used in calculating *C*1–*C*4, the bounds in equations F21.0–E.21.3 include the following data elements described in section II.C: (1) the means of the outcome measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups ($\overline{y}_{iy}$, and $\overline{y}_{cy}$); (2) the means of the outcome measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups ($\overline{y}_{ixy}$, and $\overline{y}_{cxy}$); (3) the standard deviations of the outcome measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data, which are used to calculate s_x [88]; and (4) the number of subjects with observed outcome data in the analytic sample by condition (n_i , and n_c).

Applying the bounds $C1^{*}$ – $C4^{*}$ does not require knowing the baseline effect size using imputed baseline data. Rather, these bounds use the complete case baseline effect size. When the authors impute the baseline data using an acceptable approach and the manuscript reports the baseline effect size based on imputed data, g_{xI} , a different set of bounds should be used.

Comparing g_{xMAR} and g_{xNMAR} , the bias in the imputed baseline effect size due to deviations from MAR is given by:

[88] For simplicity, this is referred to using the consistent notation despite the difference in the data used to calculate it.

[E.22] $$Bias_x = \frac{1}{s_y}\left\{\left(\frac{n_{iy}}{n_i}\right)\left(g_i(\rho)-\rho\right)\left[\overline{y}_{iy}-\overline{y}_{ixy}\right]-\left(\frac{n_{cy}}{n_c}\right)\left(g_c(\rho)-\rho\right)\left[\overline{y}_{cy}-\overline{y}_{cxy}\right]\right\}.$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size $D1^*$ – $D4^*$:

[E.23.0] $$D1^* = \omega\left|g_{xI}\right|$$

[E.23.1] $$D2^* = \omega\left|g_{xI}+\frac{1}{s_y}\left(\frac{n_{iy}}{n_i}\right)\frac{1-\rho^2}{\rho}\left[\overline{y}_{iy}-\overline{y}_{ixy}\right]\right|$$

[E.23.2] $$D3^* = \omega\left|g_{xI}-\frac{1}{s_y}\left(\frac{n_{cy}}{n_c}\right)\frac{1-\rho^2}{\rho}\left[\overline{y}_{cy}-\overline{y}_{cxy}\right]\right|$$

[E.23.3] $$D4^* = \omega\left|g_{xI}+\frac{1}{s_y}\frac{1-\rho^2}{\rho}\left[\left(\frac{n_{iy}}{n_i}\right)\left(\overline{y}_{iy}-\overline{y}_{ixy}\right)-\left(\overline{y}_{cy}-\overline{y}_{cxy}\right)\right]\right|.$$

The formulas for $C1^*$ – $C4^*$ and $D1^*$ – $D4^*$ reduce to $C1$ – $C4$ and $D1$ – $D4$ when there are no missing outcome data.

References

Allison, P.D. "Missing Data." Paper No. 136. Thousand Oaks, CA: Sage University, 2002.

Andridge, R.R., and R.J.A. "Little. Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics*, vol. 27, no. 2), 2011, pp. 153–180.

Azur, M.J., E.A. Stuart, C. Frangakis, and P.J. Leaf. "Multiple Imputation by Chained Equations: What Is It and How Does It Work." *International Journal of Methods in Psychiatric Research*, vol. 20, no.1, 2011, pp. 40–49.

Little, R.J., R. D'Agostino, M.L. Cohen, K. Dickersin, S.S. Emerson, and J.T. Farrar, et al. "The Prevention and Treatment of Missing Data in Clinical Trials." *The New England Journal of Medicine*, vol. 367, no. 14, 2012, pp.1355–1360.

Little, R.J.A., and D.B. Rubin. *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley, 2002.

Peugh, J.L., and C.K. Enders. "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement." *Review of Educational Research*, vol. 74, no. 4, 2004, pp. 525–556.

Puma, M.J., R.B. Olsen, S.H. Bell, and C. Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley, 1987.

Schafer, J.L. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research*, vol. 8, no. 1, 1999, pp. 3–15.

Wooldridge, J.M. "Econometric Analysis of Cross Section and Panel Data." Cambridge, MA: MIT Press, 2002.

This page has been left blank for double sided copying.

This page has been left blank for double sided copying.

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