MATHEMATICA
Policy Research

Learning about Infant and Toddler Early Education Services (LITES): Review Protocol

December 2015

Shannon Monahan
Jaime Thomas
Lauren Murphy
Diane Paulsell

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

CONTENTS

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

TABLES

FIGURE

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

I. CHARACTERISTICS OF PROGRAM MODELS ELIGIBLE FOR REVIEW

The Office of the Assistant Secretary for Planning and Evaluation (ASPE), in partnership with the Administration for Children and Families (ACF) within the U.S. Department of Health and Human Services, funded Mathematica Policy Research and its partners to conduct the Learning about Infant and Toddler Early Education Services (LITES) project. LITES aimed to identify effective and replicable program models to support infant and toddler early learning in out-of-home early care and education (ECE) settings to inform future research, policy, and program directions at the federal, state, and local levels.

LITES had two main components: (1) a systematic review to identify effective program models to support infant and toddler early learning in out-of-home ECE settings, and (2) a scan of the field for program models that are compelling but lack rigorous research examining impacts on children's developmental outcomes. This report accompanies the systematic review, and provides a detailed description of the methodology used for the review.

The LITES systematic review focused on program models designed to improve outcomes in language, cognition, and/or social-emotional/behavioral development for infants and toddlers.[1] To be considered eligible for inclusion in the LITES review, we required program models to meet the following criteria:

- **Replicable components with a focus on supporting early learning.** Eligible models fell into one of three categories. Direct multicomponent models provided a defined set of replicable program components, including direct early learning services to infants and toddlers in out-of-home ECE settings. Direct enhancement models had at least one replicable program component and provided direct early learning services to infants and toddlers in out-of-home ECE settings. Indirect enhancement models consisted of professional development programs with replicable program components that focused on helping adult out-of-home caregivers support infant and toddler early learning.
- **A focus on infants and toddlers.** The target population for the models had to include infants and toddlers, defined as children from birth to age 36 months, or their adult out-of-home caregivers. Models could include children from other age groups as well. For example, models could target children from birth to age 5, or the programs could begin prenatally. However, the primary focus of the models had to be on supporting infant and toddler early learning in out-of-home ECE settings.
- **Broad targeting.** Models had to be targeted broadly to infants and toddlers and/or their adult out-of-home caregivers. Models narrowly targeting infants and toddlers with diagnosed disabilities or specific medical conditions were not included in the review.[2]

[1] Appendix H contains a glossary of research terms.

[2] The federal government currently makes specific investments in special education and to support the development of children with disabilities. This review focused on identifying effective program models for supporting early learning among a broad range of infants and toddlers.

However, models targeting broad groups of at-risk infants and toddlers (for example, children from low-income families or low birth weight children) were eligible for inclusion.[3]

- **Out-of-home delivery.** Services had to be provided outside of the children's homes. Models could be implemented in center-based settings, such as child care centers, or in home-based settings, such as family child care homes or informal caregivers' homes. Program models that provided supplemental home visits were eligible for inclusion in the review, but the primary setting had to be out-of-home care. Similarly, program models that provided supplemental services in areas such as nutrition, health and developmental screening, supports for parents, and referrals to other community resources were considered for inclusion in the review. However, the primary focus of services delivered outside the child's home had to be on supporting infant and toddler early learning.
- **Specific criteria for indirect enhancement models.** Professional development programs delivered to adult out-of-home caregivers were eligible for inclusion in the review if the programs involved intervening directly with caregivers, took place in the caregiving or a similar setting, and focused on helping caregivers support infant and toddler early learning.[4]

For an eligible program model to be included in the systematic review, it had to have at least one study that met the LITES study inclusion criteria outlined next in Chapter II.

[3] Although the review targets children broadly, subgroups of particular interest include children from low-income families, dual-language learners and immigrants, children from minority racial and ethnic groups, children with special needs, and children in author-defined risk groups.

[4] Other indirect services—such as parenting, family self-sufficiency, or referral services—were not eligible for the review, because they did not target children's early learning in out-of-home care settings. However, outcomes in some of these domains—such as parenting—were recorded as part of LITES, if reported in the original studies. See Appendix A, Table A.1 for a full list of LITES outcome domains.

II. STUDY INCLUSION CRITERIA

To be considered eligible for inclusion in the LITES review, we required that studies meet the following criteria:

- **Study sample.** Study samples had to include children enrolled in the program before 36 months of age (including prenatal enrollment). If the sample contained children older than the target age range, we reported on disaggregated results for those enrolled before age 36 months, when possible. If disaggregated study results were not available, we required that 50 percent or more of the sample be younger than 30 months at the time of program enrollment.
- **Outcomes of interest.** We required that the study include at least one outcome in any of the following child outcome domains:[5]
  - **Cognitive development,** including outcomes such as attention, memory, object permanence, concept development and categorization, understanding relationships (for example, cause and effect), spatial reasoning, and problem solving
  - **Social-emotional/behavioral development,** including outcomes such as emotion regulation, impulse control, sociability, and attachment
  - **Language development,** including outcomes such as receptive language, expressive language (including gestures), joint attention, and emergent literacy skills (for example, listening comprehension)
- **Language of publication.** The study must have been published in English.
- **Publication time frame.** The study must have been published in 1960 or later.
- **Study design.** Eligible designs for review included randomized controlled trials (RCTs), matched comparison group designs (MCGDs), single case designs (SCDs), and regression discontinuity designs (RDDs).

[5] Child health outcomes, such as height, weight, and hospitalizations, were also reported if present in a study of a *model* that had at least one study with child outcomes in a cognitive, social-emotional/behavioral, or language domain.

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

III. SEARCH, SCREENING, AND IDENTIFICATION STRATEGY

We used a multistep process to search, screen, and identify studies eligible for review. First, we conducted a comprehensive search for relevant literature in databases and issued a call to researchers and stakeholders in the ECE field for relevant studies. We then screened the studies to identify citations potentially relevant to the review. From the studies that passed this screening, we identified program models for review, and then conducted a targeted search for studies of the identified models using the model name as a search term.

A. Phase I: Search for relevant literature

We searched for all relevant literature, including unpublished literature that aligned with the LITES review scope and study inclusion criteria. To maximize our search results, we implemented the following search techniques, using a four-step process:

Step 1: Develop focused search terms. Working with the Federal Project Officers, the LITES consultants, and our librarians, we developed search terms intended to capture the range of relevant literature on ECE services for infants and toddlers. The search terms captured evaluations of relevant models, and were carefully ordered and selected to ensure that a high proportion of identified citations were potentially relevant to the review. Table III.1 shows the search terms.

Table III.1. Search terms for the LITES review

Category	ID	Search term
Search Restrictions	--	Studies published in English only Studies published during or after 1980[a]
Activity	S1	[(early near educat*) or preschool or "pre-school" or childcare or "child care" or daycare or "day care" or "nursery school*" or "early learn*" or "nonparental care" or "non-parental care" or "early care" or "center based" or "center-based" or "infant care" or "toddler care" or "early childhood" or "child develop*] And (program* or intervention* or service* or model*)
Target Group	S2	"birth to three" or "zero to three" or baby or babies or infan* or newborn* or toddler* or "young child*" or (birth near "36 mo*") or (prenatal near "36 mo*") or "birth to 3" or zero to 3" or "0 to 3" or "preschool child*"
Outcomes	S3	(cognit* or language or linguistic or "social-emotional" or "socioemotional" or "socio-emotional" or "social and emotional" or behavior* or health*) and (develop* or domain* or "school readiness" or "school achievement" or "child develop*" or intelligence or IQ or "executive function*" or vocabulary or "social skills" or "self regulat*" or aggress* or attach* or "learn*" or "quality" or outcome)
Document Type	S4	(study or studies or evaluat* or research or trial or experiment* or "clinical trial" or "controlled clinical trial" or "controlled study" or "randomized control trial" or longitudinal stud* or "program evaluation") and (effect* or efficac* or impact* or outcome* or evidence or implement* or fidelity or cost* or replic* or finding* or result*)
Combine Terms	S5	S1 AND S2 AND S3 AND S4

Note: When performing proximity searches (for example, quality near child care), we used a parameter that defined "near" as "within five words" to find relevant literature without capturing a large volume of irrelevant literature. Searches looked back to 1980 (1960 in the targeted search phase described in Section D of this chapter) only if a given database had literature of that age; otherwise, we began the search at the earliest available date.

[a] During the targeted search phase, we extended the time frame to 1960.

Step 2: Database search. Using the focused search terms, the Mathematica library staff searched titles, abstracts, subjects, and keywords within numerous databases. Table III.2 lists the databases, and Appendix B describes each database. Mathematica librarians used advanced searching techniques—such as proximity searches (for example, requiring the words *early*, *childhood*, and *education* to be within five words of one another)—to optimize our ability to find relevant literature. Databases differ in how they organize content; therefore, the librarians tailored the search methods to the databases and checked the project search terms against keyword and subject terms for each database when possible to ensure that we did not overlook relevant citations. The search strategies were documented for future replication. The librarians saved literature search results in a designated project account created in RefWorks, an online (but private and password-protected) bibliographic management system that enables storing, scanning, and sorting a customized list of study citations and abstracts.

Table III.2. Databases for LITES literature search

Academic Search Premier	Campbell Collaboration
Child Care and Early Education Research Connections	CINAHL with Full Text Cochrane
Cochrane Central Register of Controlled Trials	Cochrane Database of Systematic Reviews
Cochrane Methodology Register	Database of Abstracts of Reviews of Effects
EconLit	Education Research Complete
E-Journals	ERIC
MedLine	PsycINFO
ProQuest Dissertations & Theses	SAGE Journals
SocINDEX with Full Text	Scopus

Step 3: Reference check. To ensure that the literature search was thorough and comprehensive, we compared the references in other ECE literature reviews with the results from our database searches. The review team attempted to diagnose why some studies were not initially located and conducted further searches with additional targeted search terms. We compared our results against the studies collected by the "Effects of Early Childhood Programs on Children: A Comprehensive Meta-Analysis" being conducted for the National Institute of Child Health and Human Development by Greg Duncan, Katherine Magnuson, Holly Schindler, and Hirokazu Yoshikawa. We also compared our results against the following:

Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). Early childhood interventions: Proven results, future promise. Report to the PNC Financial Services Group, Inc. Santa Monica, CA: Rand Corporation. Available at http://www.rand.org/pubs/monographs/2005 /RAND_MG341.pdf.[6]

Leak, J., Duncan, G. J., Li, W., Magnuson, K., Schindler, H., & Yoshikawa, H. (2010). Is timing everything? How early childhood education program impacts vary by starting age, program duration and time since the end of the program. Paper presented at the Biennial Meeting for the Society for Research on Child Development, Montreal, Quebec, March 31–April 2, 2011.

[6] This was used in the Home Visiting Evidence of Effectiveness (HomVEE) literature search. Hybrid models are included.

Meisels, S. J., & Shonkoff, J. P. (Eds.). (1990). *Handbook of early childhood intervention*. New York: Cambridge University Press.

Shonkoff, J. P., & Meisels, S. J. (Eds.). (2000). *Handbook of early childhood intervention* (second edition). New York: Cambridge University Press.

Step 4: Call for studies. We issued a call for studies to help find additional literature. Grey literature, including dissertations and unpublished studies, might be relevant to the LITES review, but such literature can be difficult to find. Some databases (for example, citations obtained by searching the Campbell Collaboration) and recommendations of federal staff and consultants helped us find relevant grey literature sources. In work on other evidence reviews, we have found that combining input from experts with a broad call for papers is the most successful strategy for capturing relevant unpublished work. A public call for papers also promotes transparency and engages the early childhood field in the project. The call for papers that we issued in January 2014 described our inclusion criteria to ensure that we captured relevant studies, especially those that were unpublished or under review and might not appear in the database search. The call for papers also included the purpose and background of the project, and provided instructions regarding the format, method, and deadline for sending materials to us.

To widely distribute the call for studies among researchers, program evaluators, policy experts, and other stakeholders, the project team emailed the call to a broad range of electronic mailing lists, including research and policy organizations, key early childhood professional associations and practitioner groups, and university-affiliated research centers. Table III.3 presents the distribution list (See Appendix C for the call for studies). We also sent the call for papers to our consultants and expert panel and asked them to disseminate it to their colleagues.

The project used a dedicated email address [LITES@mathematica-mpr.com] to receive and acknowledge submissions through the public call for studies. As submissions arrived, submitters received an automatic reply during the open call period, and the project team catalogued and screened each document.[7] If the citation was not already represented in our records, the team added it to the collection of research identified through database searching. We accepted submissions for an eight-week period. If submissions arrived after the call for papers closed, an automatic email reply was generated. We accepted late submissions at the discretion of the Federal Project Officers.

Table III.3. Distribution list for LITES call for studies

Group	Email or contact information
American Academy of Pediatrics	kidsdocs@aap.org
American Education Research Association	aeainfo@vanderbilt.edu
American Evaluation Association	info@eval.org
American Medical Association	mediarelations@jama-archives.org
American Professional Society on the Abuse of Children	apsac@apsac.org
American Psychiatric Nurses Association	tlantrip@apna.org
American Psychological Association	public.affairs@apa.org

[7] We created another project email account for corresponding with study authors as part of the review process.

Group	Email or contact information
American Public Health Association	comments@apha.org
American Sociological Association	publications@asanet.org
Association for Psychological Science	amikulak@psychologicalscience.org
Association of Maternal and Child Health Programs	info@amchp.org
Association for Public Policy Analysis and Management	appam-l@list.s-3.com
Child Care and Early Education Research Connections	contact@childcareresearch.org
Child Maltreatment Researchers Listserv	child-maltreatment-research-l@cornell.edu
Child Welfare Information Gateway	info@childwelfare.gov
Center for Law and Social Policy	jrobinson@clasp.org
Coalition for Evidence-Based Policy	danderson@coalition4evidence.org
Collaborative for Understanding the Pedagogy of Infant/Toddler Development	vallotto@msu.edu
Early Head Start Research Consortium	ehs_research@listserve.icfi.com
Evidence Based Home Visitation Programs	ebhv@listserve.icfi.com
Federal Inter-Agency Workgroup on Child Abuse & Neglect	catherine.nolan@acf.hhs.gov
FRIENDS Listserv for Community Based Child Abuse Prevention Grantees and Interested Community Members	friendsnrc@lists.friendsnrc.org
Foundation for Child Development	info@fcd-us.org
Future of Children	foc@princeton.edu
Harvard's Center on the Developing Child	developingchild@harvard.edu
Healthy Start Eval Listserv (NIH)	healthystarteval@list.nih.gov
HRSA Traumatic Brain Injury Technical Assistance Center Listserv	tbiserv@list.nih.gov
International Society for the Prevention of Child Abuse and Neglect	ispcan@ispcan.org
International Society on Infant Studies	lewkowic@fau.edu
Maternal and Child Health, ECCS Listserv	eccs@lists.ucdenver.edu
MCH Training Listserv Members	mchtraining@list.nih.gov
National Association for Welfare Research and Statistics	NAWRS2013@gmail.com
National Association for the Education of Young Children	membership@naeyc.org
National Association of Social Workers	membership@naswdc.org
National Council on Family Relations	info@ncfr.org
Network of Infant/Toddler Researchers	nitr@lists.icfwebservices.com
Partners in Maternal and Child Health Safety Net Listserv	Members are contacted directly
Pew Charitable Trusts	info@pewtrusts.org
Prevent Child Abuse America	mailbox@preventchildabuse.org
Prevention Subcommittee Distribution List	Members are contacted directly
Social Work Research Network (formerly called Institute for the Advancement of Social Work Research)	swrnet@bu.edu
Society for Prevention Research	info@preventionresearch.org
Society for the Psychological Study of Social Issues	spssi@spssi.org
Society for Research in Child Development	info@srcd.org
Society of Pediatric Nurses	spn@dancyamc.com
Zero to Three	0to3@presswarehouse.com

B. Phase II: Implement screening procedures

After we completed the literature search, trained staff conducted a multistep screening procedure to identify the most relevant citations:

Step 1: Preliminary screening. In this step, we removed citations from our list that were not useful to the review.

- **Deduplication of citations.** When using the search terms across multiple databases, searches sometimes identified the same citation in more than one database. We kept only one copy of each citation, deleting the others from RefWorks.

- **Exclude publications that are not studies.** Screeners next eliminated any irrelevant citations returned by the search terms (specifically, those that had our keywords but might not be studies of programs, such as letters to the editor, book reviews, or press releases). These were not considered further but remained in RefWorks labeled as nonstudies.

Step 2: Screening in RefWorks. After the removal of nonstudies, additional screening for relevance was necessary using the study abstracts. For example, when searching in medical journals, we might locate studies about how young children fare when offered a specific nutrition plan at home. This would be out of scope for the review but could be captured in the broader search. Citations screened out at this stage were retained in RefWorks but assigned a disposition code (see Appendix D, Table D.1) describing the reason for their exclusion. We screened studies for the following factors:

- **English publication.** Excluded studies not published in English.
- **Policy relevant.** Excluded studies of models delivered in a developing-world context.
- **Possible to attribute effects solely to the model of interest.** Excluded studies in which it was not possible to attribute effects solely to the model of interest. For example, studies in which a direct multicomponent model of interest was combined with another direct multicomponent intervention were excluded.
- **Published 1960 or later.** Excluded studies published before 1960.
- **Primary study.** Excluded summaries of studies reported elsewhere (for example, literature reviews or meta-analyses).
- **Target population in range.** Excluded studies in which the children or families were not enrolled in the program model before the child reached 36 months of age. To target models for children from birth to 36 months of age, we required results disaggregated for those enrolled before age 36 months. If disaggregated study results were not available, we required that 50 percent or more of the sample be younger than 30 months at the time of program enrollment.[8]
- **Services relevant to the review.** Excluded studies that were not (1) direct multicomponent models that provided a defined set of replicable program components, including direct early learning services to infants and toddlers in out-of-home ECE settings; (2) direct enhancement models with at least one replicable program component and provided direct early learning services to infants and toddlers in out-of-home ECE settings; or (3) indirect enhancement models consisting of professional development programs with replicable program components that focused on helping adult out-of-home caregivers support infant and toddler early learning. In addition, excluded studies in which services primarily targeted children with specific disabilities or medical conditions.

[8] This criterion is similar to several What Works Clearinghouse (WWC) review protocols that use a 50 percent threshold for defining eligible study samples when results are aggregated (such as the Early Childhood Education for Children with a Disability topic area protocol). To exclude ECE services that focused primarily on children 36 months and older, we set the threshold at 30 months.

- **Primary service delivery location is out of the home.** Excluded studies of models in which out-of-home ECE services were not the primary service delivery mechanism (for example, those that primarily delivered services through home visits).[9]
- **Replicable program model.** Excluded studies in which the ECE services under study did not include a defined package of replicable program components.
- **Subgroups out of scope.** Excluded studies that only reported on subgroups that were not the LITES pre-identified subgroups of interest.
- **Eligible outcomes.** Excluded studies that did not measure at least one child outcome in one of the following domains: cognitive, language, or social-emotional/behavioral development.
- **Eligible design.** Excluded studies that did not use one of the eligible designs: RCTs, MCGDs, SCDs, or RDDs. We coded ineligible designs in the database to retain supplemental information about the models prioritized for the systematic review.[10]

Step 3: Screening in SharePoint. When we identified relevant citations and those that required more information, we transferred them to a secure project-specific Microsoft SharePoint website, through which the team could store information about each citation, link to the full text of studies, and upload completed reviews. The SharePoint site also made real-time monitoring of the screening and review progress easier. Screening continued in SharePoint, as needed, using the preceding criteria. Citations screened out at this stage were retained in SharePoint.

Step 4: Design screening. In this step, we coded the study design. We coded ineligible designs in the database to retain for the compelling-models review[11] or to supplement information about the models identified for the systematic review.[12]

C. Phase III: Identify models for review

After the search and screening phases, we analyzed the remaining citations and identified a list of replicable program models with eligible studies for review. We sorted the resulting list of models into three categories of program models. The first included direct multicomponent models that provided out-of-home early learning services for infants and toddlers. The second included direct enhancement models that could be layered on another model and typically focused on improving child outcomes in a single domain. The third included indirect enhancement models that could be layered on another model and focused on improving caregiver practice.

[9] Research on hybrid models (such as models that include both home visiting and center-based components) could be included if out-of-home services were the primary service delivery mechanism.

[10] Studies of process, fidelity, cost, sustainability, and implementation, as well as correlational, descriptive, pre-post design, and ethnographic studies, were not eligible for review because these study designs did not allow a researcher to confidently determine that the intervention under study caused changes observed in children's outcomes.

[11] See Del Grosso et al. (2015) for a description of compelling models.

[12] Other types of literature include studies of process, fidelity, cost, sustainability, and implementation, as well as correlational, descriptive, pre-post design, and ethnographic studies.

D. Phase IV: Targeted search on selected models

After we identified replicable program models, we repeated the search and screening process to locate additional literature specific to them. We included the model names as key search terms in the database search from Phase I. We searched the full text of articles for model names when possible and extended the time frame for the targeted search to 1960. We then repeated Phase II to screen the new set of studies and check the previous set of studies excluded because of their publication dates.

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

IV. ASSESSMENT OF METHODOLOGICAL QUALITY

To evaluate studies consistently and objectively, we used criteria to assess the quality of the studies prioritized for review. To maximize efficiency and build on the strong foundation established by existing evidence reviews, we began the criteria development process by reviewing existing standards from the WWC (2011; 2014), the HomVEE (2014) review, and the U.S. Department of Health and Human Services' Teen Pregnancy Prevention Evidence Review (TPP Evidence Review 2012). We tailored the WWC, HomVEE, and TPP Evidence Review standards to ensure they took into account, and were appropriate for, the distinct features and needs of the infant-toddler ECE research field. In this section, we define the four study designs eligible for inclusion in the review and describe our criteria for assessing study quality and assigning study ratings.

A. Description of eligible study designs

Eligible designs for the LITES review were RCTs, MCGDs, SCDs, and RDDs. RCTs use random assignment to create two or more groups that are, on average, similar to each other at the onset of the study (that is, at baseline).[13] These studies provide strong evidence that differences in the outcomes between the intervention and comparison groups after the implementation of an intervention (that is, at follow-up) can be attributed to the intervention, rather than to preexisting differences between the groups (Shadish et al., 2002).

In MCGDs, participants are sorted into groups through a process other than random assignment. Even if the treatment and comparison groups are well matched based on observed characteristics, they may still differ on unmeasured characteristics. Therefore, it is impossible to rule out that the findings could be attributable to unmeasured group differences.

In an SCD, each case provides its own control for comparison (WWC, 2011; 2014). A case may be a single participant or a cluster of participants (for example, a small group or classroom). For each case, the outcome variable is measured repeatedly within and across "baseline" and "intervention" phases (as they are most commonly called). Measurements taken during baseline phases, in which no intervention is applied, are compared to measurements from intervention phases, in which researchers apply the intervention to the case under study. Consistent differences between outcome measurements in the baseline and intervention phases provide evidence of an intervention's effect. SCDs can provide a strong basis for establishing causal inference, and these designs are widely used in applied and clinical disciplines in psychology and education, such as school psychology and special education.

RDDs are applicable when a continuous "scoring" rule is used to assign study units (for example, children, classrooms, or child care centers) to an intervention (WWC, 2011; 2014). Units with scores below a preset cutoff value are assigned to the intervention group, and units with scores above the cutoff value are assigned to the comparison group (or vice versa). For example, children may be assigned to a language intervention if they score below a preset point

[13] If random assignment is applied appropriately, then there are no *systematic* differences between the two groups at baseline; however, there may be chance differences. Chance differences may be more likely with small sample sizes.

on a standardized test, or child care centers may be awarded a grant based on a certain score on their application. Units close to one another on either side of the cutoff are likely to be very similar, differing only in that some were assigned to the intervention and some were not. Therefore, comparing outcomes between these two groups can give an unbiased estimate of the intervention's effect if certain conditions are met. RDDs are increasingly used by researchers to examine the effects of education-related interventions.

Study designs that lack a comparison group or condition (for example, pre-post designs) offer no way to assess what participants' outcomes would have been in the absence of the intervention. These study designs cannot rule out the possibility that changes were caused by other factors—for example, history (an event besides the intervention that could have produced the observed outcome) or maturation (participants' natural changes over time that could have produced the observed outcome) (Shadish et al., 2002). Therefore, designs lacking a comparison group or condition were not eligible for review.

B. Criteria for assessing study quality and assigning study ratings

The study quality standards focused on internal validity—that is, a study's ability to isolate the effects of a program or intervention from other factors that may influence participants' outcomes. Following HomVEE and the TPP Evidence Review, we used three study-level ratings: high, moderate, and low (HomVEE, 2014; TPP Evidence Review, 2012). The three study-level ratings provided an assessment of a study's internal validity. In brief, the high rating was reserved for RCTs with low attrition of sample members and no reassignment of sample members after the original random assignment, as well as for SCDs and RDDs that met WWC design standards without reservations (Table IV.1). The moderate rating applied to RCTs that, due to flaws in the study design or analysis (for example, reassignment of sample members), did not meet all the criteria for the high rating; MCGDs that demonstrated baseline equivalence and applied statistical controls; and SCDs and RDDs that met WWC design standards with reservations. Low-rated studies did not meet the requirements for a high or moderate rating.

Table IV.1. Summary of study rating criteria for the LITES review

LITES study rating	RCTs	MCGDs	SCDs[a]	RDDs[a]
High	Random assignment Low attrition No reassignment No confounding factors	Not applicable	Timing of intervention is systematically manipulated. Outcomes meet WWC standards for interassessor agreement. At least three attempts to demonstrate an effect. At least five data points in relevant phases.	Integrity of forcing variable is maintained institutionally AND statistically. Meets WWC attrition standards. Continuous relationship between outcome and forcing variable. Satisfies all WWC criteria for functional form and bandwidth.
Moderate	If there is reassignment or high attrition, highest possible rating is moderate and MCGD rating criteria apply.	Baseline equivalence established on required measures Proper statistical controls used No confounding factors	Timing of intervention is systematically manipulated. Outcomes meet WWC standards for interassessor agreement. At least three attempts to demonstrate an effect. At least three data points in relevant phases.	Integrity of forcing variable is maintained institutionally OR statistically. Meets WWC attrition standards. Continuous relationship between outcome and forcing variable. Satisfies selected WWC criteria for functional form and bandwidth.
Low	Studies that do not meet the requirements for a high or moderate rating			

[a] WWC SCD and RDD standards are pilot standards applied to judge evidence from individual studies. The LITES study quality criteria for SCDs and RDDs are the same as the WWC SCD and RDD standards (WWC 2011; 2014). We have made no modifications for the LITES review.

1. Threats to internal validity

Because all the studies we reviewed for LITES were RCTs or MCGDs, and we directly adopted the WWC SCD and RDD standards, we focus the remainder of this chapter on RCTs and MCGDs. In this section, we discuss the following threats to internal validity: confounding factors, attrition, and nonexperimental study designs.

Confounding factors. Confounding factors, or "confounds," threaten the internal validity of RCTs and MCGDs because, if a confounding factor is present, a study cannot distinguish between the effect of that factor and the intervention of interest. A confounding factor is often defined as a third variable related to both the independent variable and dependent variable, and that might account for the observed relationship between the two. In many cases, this occurs when some aspect of the design lines up exactly with either the intervention or comparison group. For example, if there is only one classroom in the intervention group, intervention effects are indistinguishable from classroom effects—that is, it is impossible to determine whether the intervention or another feature of the classroom, such as the teacher or the composition of the students, caused the observed outcomes.

Attrition. In the context of rating study quality, attrition is problematic in RCTs because, although randomization results in intervention and comparison groups that are similar at baseline, attrition may compromise the initial equivalence of the groups and lead to biased estimates of intervention impacts.[14] Both overall and differential attrition can contribute to bias in the estimated effect. To illustrate overall and differential attrition, consider a hypothetical study that randomly assigned 100 children to the intervention group and 100 to the comparison group. Suppose that, at the end of the intervention, 80 children remained in the intervention group and 70 remained in the comparison group. In this example, the overall attrition rate would be equal to the total number of children who left the study divided by the total number of children randomly assigned: 50/200, or 25 percent. The differential attrition rate is the absolute value of the difference between the attrition rates in the intervention and comparison groups: |20/100 – 30/100|, or 10 percent.

Nonexperimental study designs. In experimental studies, or RCTs, treatment assignment is random, which, as mentioned previously, ensures that intervention and comparison groups are similar at baseline in observable characteristics, such as socioeconomic status (SES), as well as unobservable characteristics, such as motivation to participate in the intervention. In nonexperimental designs such as MCGDs, group assignment is nonrandom, and we cannot rule out the possibility that groups differ in unobservable ways at baseline. Unobservable baseline differences can bias estimates of the intervention's impact. For example, if the intervention group contained families who, before the intervention, provided more developmental materials for their children at home than families in the comparison group (and if this difference was not controlled for in impact analyses), researchers might find cognitive development impacts that appear favorable to the intervention but are instead due to this preexisting difference between the study groups.

2. Standards to address threats to internal validity

In this section, we discuss the LITES standards designed to address the threats to internal validity mentioned previously. Appendix E contains decision trees illustrating these standards.

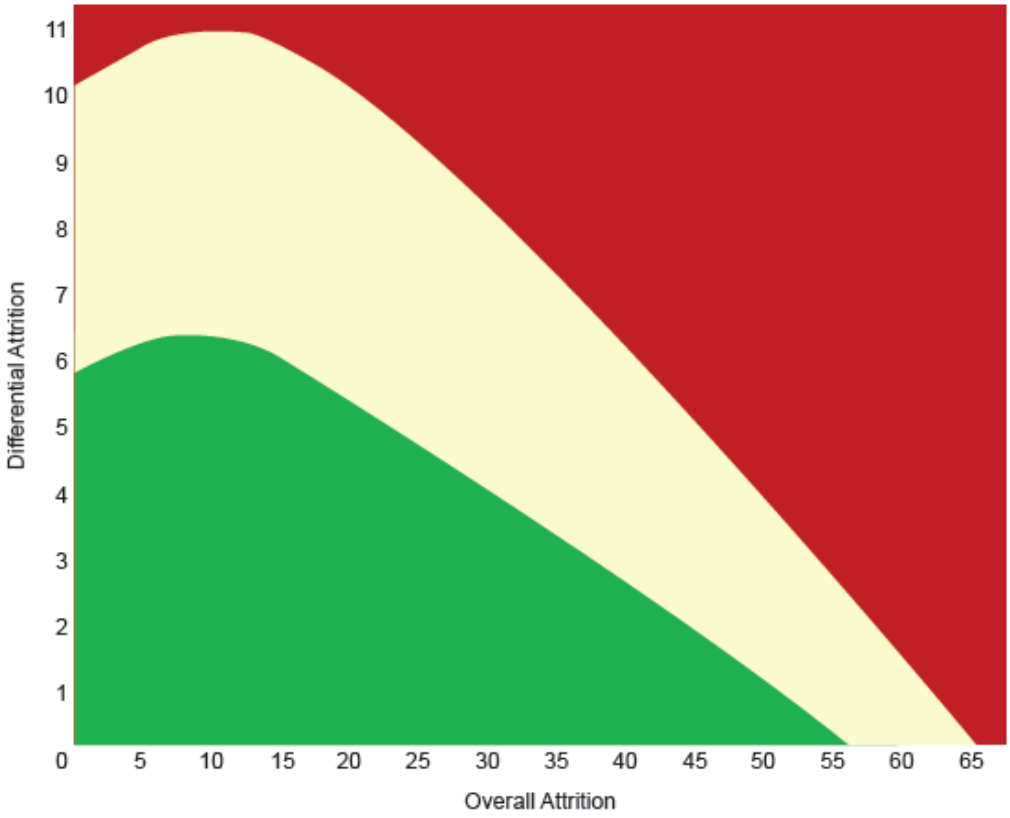
Studies with confounds received low ratings. For this review, a low rating was assigned to RCTs or MCGDs with only one unit in the intervention and/or comparison condition or other confounding factors, such as systematic differences in data collection procedures between the intervention and comparison groups.

Attrition standards set cutoffs for acceptable levels of attrition. The cutoff for an acceptable level of sample attrition is tied to the extent of overall or differential attrition and to a combination of the two (Figure IV.1). In LITES, RCTs with combinations of overall and differential attrition that fall into the green area of Figure II.1 had low attrition. RCTs with combinations of overall and differential attrition that fall into the tan or red areas had high

[14] Attrition is not a factor examined for MCGDs because for this study design type, only the *analytic* sample is considered when determining study quality.

attrition and were reviewed as MCGDs.[15] The highest possible rating for RCTs with high attrition was moderate.[16]

Figure IV.1. Standard for assessing sample attrition in study quality ratings



Note: This figure illustrates WWC attrition bounds. The WWC defined two attrition bounds: "liberal" and "conservative." The conservative bounds apply when there is reason to believe that attrition from a program is related to the intervention implemented. LITES applied the conservative attrition bounds from the WWC because attrition from out-of-home ECE interventions is plausibly related to treatment status (for example, families may leave a program that requires a high level of family engagement). The green/bottom-left region shows combinations of overall and differential attrition that yield low levels of attrition bias according to the conservative attrition bounds. The liberal attrition bounds include the green and the tan regions. The red/top-right region shows combinations that yield high levels of attrition bias in all cases.

Source: What Works Clearinghouse. *Procedures and Standards Handbook, Version 3.0.* Washington, DC: U.S. Department of Education, 2011.

Following WWC standards for clustered RCTs, in which clusters (such as child care centers) are randomly assigned to the intervention or comparison group and outcomes are assessed at the individual level, attrition was assessed at the cluster level and the individual level. Attrition had

[15] Because there is reason to believe that attrition from out-of-home ECE interventions is related to treatment status, LITES applied the conservative attrition bounds from WWC.

[16] The attrition standards do not apply to matched comparison group studies. These studies were evaluated on the basis of the final analysis sample, from which there is no attrition.

to be low at both levels to receive a high rating. If attrition was high at either level (or both levels), then the study was reviewed using the same criteria as an MCGD, and the highest possible rating was moderate (Table IV.2).

Table IV.2. Attrition standards for cluster randomized trials

Level of sample attrition		
Cluster level	**Individual level**	**Highest possible study rating**
High	Low	Moderate, with evidence of baseline equivalence and statistical controls
High	High	Moderate, with evidence of baseline equivalence and statistical controls
Low	Low	High
Low	High	Moderate, with evidence of baseline equivalence and statistical controls

Cluster correction was required. In a clustered RCT, the unit of assignment is different from the unit of analysis. For example, classrooms could be assigned to intervention and comparison conditions, but the researcher may analyze child-level outcomes. In these cases, the analysis must account for clustering (HomVEE, 2014; WWC, 2011; 2014). If a correction is not made, the statistical significance of the findings may be overstated. That is, a finding may be misclassified as statistically significant when, if clustering were properly taken into account, the finding would not be significant. If the authors did not correct for clustering at the unit of assignment, LITES made an adjustment, if sufficient information was available.

Establishing baseline equivalence was required. For MCGDs and RCTs with high attrition or reassignment, baseline equivalence of intervention and comparison groups is a key concern. Demonstrating baseline equivalence means showing that the intervention and comparison groups have similar observable characteristics at baseline. This supports conclusions that the intervention—rather than preexisting differences—led to the observed outcome (Shadish et al., 2002). For this review, equivalence had to be established on the final analytic sample used in the analysis of follow-up outcomes (not the baseline sample). It is important to establish baseline equivalence on key variables rather than merely adjusting for these variables by including them as covariates in a regression, because establishing baseline equivalence provides some assurance that intervention and comparison groups overlap enough with respect to these characteristics to enable a reasonable estimation of the program effect. If there is little overlap, the regression-based approach depends heavily on the model's functional form assumptions—that is, how accurately the model captures the true relationship between the covariates and the outcome. In this case, impact estimates rely heavily on extrapolation (Stuart, 2010), and such extrapolations can be highly sensitive to functional form (Foster, 2003).

In LITES, baseline equivalence was established if there were no statistically significant differences on specified variables (described below) for the analytic sample at baseline.[17] The

[17] Variables upon which baseline equivalence must be established vary by evidence review, but typically include demographic information and pre-intervention outcomes. LITES required studies to establish baseline equivalence on demographic characteristics but not child outcome measures. Child outcome measures were not

LITES review used author-reported baseline equivalence calculations, if available, and preferred two-tailed tests with $\alpha = 0.05$. When necessary, the LITES team calculated baseline equivalence and used a *p*-value from a chi-squared test for categorical variables (including dichotomous variables).

LITES required that baseline equivalence be established on:

- Race/ethnicity
- SES
- Child age

Demographic variables such as race/ethnicity and SES are commonly available and have been shown to be related to outcomes of interest. For example, research demonstrates links between SES and outcomes such as child health and child cognitive and social-emotional development (Bradley & Corwyn, 2002). SES can be measured in multiple ways, but we preferred equivalence on maternal education, income, earnings, or poverty levels according to federal thresholds. We also considered alternative measures of SES (that is, employment and Aid to Families with Dependent Children [AFDC] and Temporary Assistance for Needy Families [TANF] or food stamps receipt), if at least two such alternative measures of SES were provided. Assessment of age is important in predicting cognitive and social-emotional development outcomes, even in models that include race/ethnicity and multiple measures of SES.[18]

In addition to these variables, a study may present comparisons for other factors at baseline that might predict later outcomes, such as family structure, maternal behaviors, birth weight, or age at which early developmental milestones were attained. If any variables collected at baseline were not equivalent, the study may have been downgraded (that is, no longer eligible to receive the highest rating for its design). The decision to downgrade depended on the magnitude of these differences and the variables under consideration.

Statistical controls were required. In addition to establishing baseline equivalence, we required that MCGDs and RCTs with high attrition or reassignment did at least one of the following:[19]

required because, for infants and toddlers, these measures are not necessarily predictive of future outcomes, and the same measures are not always available for assessment at baseline and follow-up (for example, if a family enrolls in a study prenatally, there will not be child outcome baseline variables).

[18] The LITES review preferred that authors provide statistical evidence that groups were not significantly different in child age at assessment. However, because many outcomes were assessed within narrow age ranges (for example, the six-month Bayley measure of infant development), the review accepted credible author assertions that children in different study groups were assessed at the same age as a proxy for statistical evidence of age equivalence.

[19] Although including statistical controls (such as pre-tests or sociodemographic characteristics) can improve the precision of impact estimates (Deke et al., 2010), the LITES review did not require statistical controls or covariate adjustment for RCTs with low attrition and no reassignment.

- Use some type of covariate adjustment when estimating impacts. To meet this requirement, a study could control for any or all of the required baseline characteristics (that is, race/ethnicity, SES, and/or age) or use different controls that could help reduce bias.[20]
- Demonstrate that results are not sensitive to the statistical controls used. For example, a study could present a table of results from different models that included different sets of control variables, or state that impacts were estimated using models with different control variables but results were similar in sign, magnitude, and significance levels, regardless of model.

3. Outcome- and study-level ratings

Outcomes within a study often receive different ratings. For example, some outcomes in an RCT might have low attrition and receive a high rating. Other outcomes might have high attrition and could receive a moderate or low rating, depending on whether baseline equivalence was established and proper statistical controls were used.

Taking into account the possibility that outcomes within a study receive different ratings, LITES reported study-level ratings as follows:

- **High:** The study had at least one high-rated outcome.
- **Moderate:** The study had at least one moderate-rated outcome and no high-rated outcomes.
- **Low:** The study had no moderate- or high-rated outcomes.

C. Assessing evidence of effectiveness

In consultation with ASPE, ACF, and an expert work group, we adapted criteria for assessing evidence of effectiveness from the WWC, the HomVEE review, and the TPP Evidence Review. The LITES team customized these criteria for evaluations of out-of-home early learning programs for infants and toddlers. We examined eligible outcomes from all high- and moderate-rated studies to determine the strength of the evidence of effectiveness for each program model. All child outcomes within the cognitive, social-emotional/behavioral, language development, and child health domains that met our criteria for a high or moderate rating were deemed eligible to provide credible evidence of program effects.[21]

We also recorded information on outcomes in children's long-term risk and economic well-being domains (for example, cigarette use at age 30) and in interim outcome domains (for example, parent- or caregiver-child interaction), but these outcomes did not influence a program model's evidence of effectiveness rating. Appendix A, Table A.1 contains the primary child, children's long-term risk and economic well-being, and interim outcome domains reported in LITES.

[20] Studies use a wide variety of control variables. If, for example, a study established baseline equivalence on all required variables but used other important variables as controls, we would not downgrade it. Endogenous covariates, or variables that were assessed after baseline and might have been influenced by the intervention, were not eligible to be used as control variables.

[21] When a study followed participants from childhood through adolescence or adulthood, we continued to consider outcomes within these domains as eligible to provide evidence of effectiveness.

1. Extracting and documenting data

We extracted basic information on all outcomes in the primary child, children's long-term risk and economic well-being, and interim outcome domains that were reported in a study. For outcomes rated high or moderate, we recorded the impact estimates reported by authors and whether the impacts were favorable, unfavorable, or neutral to the intervention.[22] We also recorded the statistical significance of the impact estimates and their effect sizes or the information necessary to calculate them, when the information was available.[23]

The review team documented all this information as the study reported it, including composite, scale-level scores and subscale scores of a measure, if reported separately. We based the evidence of effectiveness rating on subscales when they were the only measures available and on composite, scale-level measures when they were the only measures available. When both types were available, we based the evidence of effectiveness rating on subscales and composite measures, as long as the composite measure provided additional information beyond that contained in the subscales. If the composite measure overlapped entirely with the subscales, we reported the subscales only.

2. Assessing the evidence of effectiveness of individual program models

Based on the information about eligible outcomes, the review team assessed the extent of evidence for each program model.[24] We assigned one of four domain-specific evidence of effectiveness ratings for each of the primary child outcome domains (cognitive, language, or social-emotional/behavioral development) and child health, if reported:

- Favorable effects: Evidence of a favorable effect with no overriding contrary evidence
- Mixed effects: Evidence of inconsistent effects
- No discernible effects: No affirmative evidence of effects
- Unfavorable effects: Evidence of an unfavorable effect with no overriding contrary evidence

LITES defined favorable and unfavorable effects as those that were statistically significant ($p \leq 0.05$) or that had an effect size greater than or equal to 0.2 standard deviations in absolute value. That is, results satisfying either of these two criteria counted toward an evidence of effectiveness rating. This decision was made because small studies would be less likely than large studies to demonstrate significant effects since smaller sample sizes are associated with

[22] An impact estimate with a positive sign is not necessarily favorable—for example, measures of problem behaviors.

[23] We recorded information on magnitudes and standard errors as presented by study authors. If authors did not report effect sizes, LITES attempted to compute them in a uniform manner (using Hedges' *g*, as in the WWC) when the necessary information was available (namely, intervention and comparison group outcome measure means, standard deviations, and sample sizes).

[24] We used categorizations similar to those of the evidence of effectiveness ratings developed by the WWC but tailored the terminology for the LITES literature. For example, the WWC refers to positive and negative effects; LITES uses "favorable" and "unfavorable."

larger *p*-values. Therefore, if statistical significance had been the only criterion for demonstrating an effect, there would have been a bias towards studies with larger sample sizes.

We applied these ratings to end of intervention outcomes and to sustained or delayed outcomes—that is, outcomes measured one year or more after the end of the intervention.[25] We did not apply any multiple comparisons corrections when assessing domain-specific evidence of effectiveness.[26] Table IV.3 provides an overview of these ratings.

Based on the domain-specific ratings, we assessed whether a program model exhibited evidence of effectiveness. If a program model exhibited favorable effects for end of intervention or sustained or delayed outcomes within any of the four primary child outcome domains, we deemed that model as exhibiting evidence of effectiveness.

Table IV.3. LITES evidence of effectiveness ratings

Domain rating	Outcome evaluation criteria
Favorable effects: evidence of a favorable effect with no overriding contrary evidence	At least one high- or moderate-rated study shows at least one significant or substantial favorable effect,[a] AND No high- or moderate-rated study shows any significant or substantial unfavorable effects
Mixed effects: evidence of inconsistent effects	At least one high-or moderate-rated study shows at least one significant or substantial favorable effect, AND At least one high- or moderate-rated study shows at least one significant or substantial unfavorable effect
No discernible effects: no affirmative evidence of effects	No study shows any significant or substantial effects, either favorable or unfavorable
Unfavorable effects: evidence of an unfavorable effect with no overriding contrary evidence	At least one high- or moderate-rated study shows at least one significant or substantial unfavorable effect, AND No high- or moderate-rated study shows any significant or substantial favorable effects

[a]A **significant effect** is statistically significant ($p \leq 0.05$). A **substantial effect** has an effect size greater than or equal to 0.2 standard deviations in absolute value.

3. Reporting of subgroup-specific outcomes

Outcomes reported for a study's full sample contributed to a program model's overall evidence of effectiveness rating. We also rated a program model's subgroup-specific evidence of

[25] End of intervention outcomes included those measured at 36 months and/or those measured at the end of the intervention. These ratings would also apply to replicated outcomes—that is, outcomes measured in two or more non-overlapping study samples—but none of the reviewed program models had any replicated effects.

[26] Mathematica's experience conducting the HomVEE systematic review taught us that authors do not commonly provide all of the information necessary to make multiple comparisons adjustments (namely, exact *p*-values). To avoid overburdening study authors with excessive author queries, we chose not to query them for this information. To provide some indication of whether a significant effect was due to chance, we report the number of significant effects as well as the number of null effects for each outcome domain.

effectiveness on outcomes in the domains of interest to LITES if they were reported separately. Subgroups of particular interest for this review were:[27]

- Children from low-income families
- Dual-language learners and/or immigrants
- Children from minority racial and ethnic groups
- Children with special needs
- Children in study-defined risk groups

[27] The full sample of a study might coincide with one of these subgroups of interest—for example, if a program model targets low-income families, a study's sample might consist entirely of low-income children. In this case, the results for the full sample would contribute to the program model's overall evidence of effectiveness rating, and to the program model's effectiveness rating for the low-income subgroup. If a study presented results for a broad sample—for example, children from low-, middle-, and high-income families—and for the subgroup of low-income children separately, the results reported for the full sample would contribute to the program model's overall evidence of effectiveness rating, and the results reported separately for the low-income subgroup would contribute to the program model's low-income subgroup effectiveness rating. Not all subgroups of interest were in the eligible studies.

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

V. REVIEW PROCEDURES

A. Conflict-of-interest policy

We had a conflict-of-interest policy to ensure separation of all staff, subcontractors, and consultants from interests in the materials under review. All Mathematica and subcontractor staff and consultants involved in the project were asked to (1) sign a conflict-of-interest statement in which they certified that they had no financial, contractual, organizational, or personal interest that would impinge on their ability to render impartial, technically sound, and objective assistance and analysis; and (2) disclose all ties to any study or model included in the review. The project director was responsible for assembling signed conflict-of-interest forms for all project team members and for monitoring for possible conflicts. To help assess possible conflicts, all parties were asked to review the following instructions and list of questions and to document any positive responses:

- Have you ever conducted research on the infant-toddler early learning program that is the subject of the study under review?
- Have you ever worked on a research project or study with one or more of the study authors?
- Have you ever worked on a research project or study with the developer(s) of the program that is the subject of the study under review?
- Do you (or any of your immediate family) have any financial interest in the infant-toddler early learning program under review? For example:
    - Are you or a family member employed by the program?
    - Are you or a family member working for an infant-toddler early learning program in your state or community?
    - Are you or a family member working for an organization that may receive funds from the program?

In addition, for transparency, we were able to, in consultation with the Contracting Officer's Representative, disclose any potential conflicts (such as instances in which a member of the project team was involved in a study reviewed by LITES) in project reports and describe how the potential conflicts were mitigated.

No Mathematica or subcontractor staff member or consultant was involved in reviewing studies in which he or she played any role, and no Mathematica staff participated in reviews of studies conducted by Mathematica. For LITES, we contracted with outside consultants to conduct first and second reviews of any studies that were conducted by Mathematica, such as the Early Head Start Research and Evaluation Project. The availability of two external reviewers ensured that the first and second reviews could be conducted by staff who were not affiliated with Mathematica for any study in which the organization had a potential conflict of interest.

In addition to maintaining a conflict-of-interest policy, asking all team members to disclose any potential conflicts of interest, and using outside reviewers to conduct reviews of studies in which Mathematica played a role, this review protocol and these documentation procedures provided further protection against the appearance of a conflict of interest. To be credible to the

field and the public, evidence reviews must be transparent, systematic, and replicable. To achieve these goals, we developed this methodical review protocol to ensure transparency. In addition to documenting the review protocol, we also documented the review findings and decisions that contributed to each study rating. Thorough documentation of each study reviewed ensures that other researchers can examine the decisions made by the reviewer and the rationale for the rating. If an author or model developer questions study ratings, the study record provides transparent and thorough documentation of our review decisions. Together with the review protocol, this documentation enables others to replicate our reviews and provides strong evidence of the review's objectivity.

B. Reviewer qualifications

The review team comprised highly qualified researchers who were certified WWC reviewers. All the reviewers have advanced degrees (including Ph.D.s, Ed.D.s, and master's degrees) in relevant fields and have extensive experience on other systematic reviews, such as the WWC and HomVEE. Most of the reviewers were from Mathematica, but, as discussed previously, two external reviewers conducted reviews of studies conducted by Mathematica.

WWC-certified reviewers attended an in-person training session led by the WWC, passed a multiple-choice test covering WWC concepts, and successfully completed a review of a study against WWC evidence standards using the WWC study review guide.

C. Reviewer training

All reviewers participated in rigorous training specific to this review, led by review task leaders. The training lasted approximately 2.5 hours. Appendix F contains the agenda. Reviewers received a course manual in advance. The training included the following topics:

- **Introduction.** We introduced the project team, discussed team members' roles and responsibilities, and thoroughly described the nature and the goals of the LITES review.
- **The review process.** The training covered the sequential review process described below, how reviewer assignments were made, the time frame for a typical review, and the correct contact person for any questions about the review or the process.
- **LITES criteria to assess study quality and evidence of effectiveness.** We described in detail the criteria for assessing study quality and evidence of effectiveness, highlighting the differences between LITES and other reviews. We also explained the LITES author query procedure (discussed below).
- **Documenting the review.** We described every aspect of completing a study review. The key element of an individual study review was a study-specific review guide, in which reviewers recorded basic study information (for example, the study citation), confirmed that the study was screened correctly, and inputted information from the study to determine its rating. During training, we introduced reviewers to the LITES study review guide and explained how they should complete each field.

D. Conducting reviews

LITES employed a rigorous, two-stage review procedure. The first stage involved two sequential reviews to assess the quality of individual studies. The second involved applying evidence of effectiveness criteria to assign an effectiveness rating based on all studies of a program model.

Two reviewers assessed a study's quality. The first reviewer evaluated the study, assigned a rating, and provided a detailed record of the study by completing a study review guide, as described previously. The second reviewer examined the study and the results of the first review. If the second reviewer disagreed with any of the first reviewer's decisions, the two reviewers discussed the differences to reach a consensus rating. An experienced reconciler confirmed all consensus rating decisions. After the rating was confirmed, the second reviewer prepared the master study review guide, which received a strict quality assurance review, described below.

After all the studies of a program model were reviewed, we used data from the study-specific review guides to apply the evidence of effectiveness criteria to assign an effectiveness rating and produce program model-specific summaries. The effectiveness ratings and the program model summaries also received a quality assurance review, as described below.

E. Quality assurance plan

After the two sequential reviews of a study were completed, the rating confirmed, and the master study review guide created, team leadership provided the quality assurance review for the guide, evaluating it for:

- Accuracy
- Consistency
- Completeness
- Clarity

Team leadership also provided quality assurance for the program model evidence of effectiveness rating and program model summary. This quality assurance review evaluated:

- The accuracy of the evidence of effectiveness rating
- The completeness and accuracy of the program model summary

F. Protocol for conducting author queries

Studies can be missing information necessary to determine their rating. Information on sample sizes at each wave of the study and on baseline equivalence of intervention and comparison groups is essential to determine a study rating. If the study noted that the author(s) conducted baseline equivalence analyses but did not report the results or provide sufficient information for LITES reviewers to assess baseline equivalence, we queried the authors for this information. If no mention of baseline equivalence analyses was made, we did not conduct an author query, and we assumed that the groups were not equivalent at baseline. This is because

the purpose of author queries was to seek clarification on existing analyses, not to suggest that study authors perform new analyses.[28]

Information on statistical significance and whether the result is favorable or unfavorable is essential to assessing effectiveness. If this information was missing, the LITES review queried authors to request the missing information.[29]

To maximize the probability of a response, we based the LITES author query protocols and procedures on those that Mathematica had developed for existing evidence reviews. Occasionally, particularly for more dated studies (such as those from the early 1980s), authors did not respond. If we did not hear from authors within a reasonable time, we assigned a rating based on available information. Appendix G contains the LITES Author Query Template that we customized for each query we sent. Table V.1 summarizes the circumstances under which the LITES team sent an author query.[30]

Table V.1. Summary of author query approach

Reasons LITES will query an author
- To obtain missing sample size information for baseline or follow-up for the analytic sample
- To obtain missing baseline equivalence information for the analytic sample
- To clarify information related to the favorability of results
- To clarify information related to the statistical significance of results
- To clarify information related to confounds, if necessary
- To clarify information related to clusters, if necessary
-To clarify information related to statistical controls, if necessary
Additional information that might be sought if a query is already being conducted
- To clarify information about study design (for example, unclear if group assignment was random)
- To obtain results disaggregated by age (if authors mention this analysis exists)
- To obtain point estimate magnitudes, effect sizes, or information to calculate effect sizes (for example, means and standard deviations)

[28] As in the WWC, HomVEE, and the TPP Evidence Review, we did not ask authors to perform new analyses via an author query.

[29] When reviewers were unable to calculate effect sizes using available information and when an author query was needed to ascertain essential information, reviewers asked authors for point estimate magnitudes, standard deviations, and/or effect sizes.

[30] If studies before 1980 were missing extensive information, project leadership and ASPE may have chosen not to send an author query.

REFERENCES

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*(1), 371–399.

Deke, J., Dragoset, L., & Moore, R. (2010). *Precision gains from publicly available school proficiency measures compared to study-collected test scores in education cluster-randomized trials.* (NCEE 2010-4003). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Del Grosso, P., Jones, C., Paulsell, D., & Monahan. S. (2015). *Learning About Infant and Toddler Early Education Services (LITES): Summarizing the research and gaps on compelling models.* Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care, 41*(10), 1183–1192.

Home Visiting Evidence of Effectiveness. (2014). *HomVEE review protocol.* Retrieved March 2014 from http://homvee.acf.hhs.gov/document.aspx?rid=4&sid=19&mid=5.

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Teen Pregnancy Prevention Evidence Review. (2012). *Identifying programs that impact teen pregnancy, sexually transmitted infections, and associated sexual risk behaviors review protocol, version 2.0.* Retrieved March 2014 from http://www.hhs.gov/ash/oah/oah-initiatives/teen_pregnancy/db/eb-programs-review-v2.pdf.

What Works Clearinghouse. (2011). *What Works Clearinghouse procedures and standards handbook, version 2.1.* Retrieved March 2014 from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf.

What Works Clearinghouse. (2014). *What Works Clearinghouse procedures and standards handbook, version 3.0.* Retrieved April 2014 from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

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

APPENDIX A

LITES OUTCOME DOMAINS

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

Table A.1. LITES outcome domains

Outcome domain	Examples
Primary child outcomes	
Cognitive development	Attention, memory, object permanence, concept development, categorization, understanding relationships (for example, cause and effect, part to whole), visual-motor integration, spatial reasoning, representational play, and problem solving
Social-emotional/behavioral development	Emotion regulation, impulse control, sociability, empathy, social problem solving, peer interaction, attachment, and adaptive behaviors (for example, self-help skills)
Language development	Receptive language, expressive language (including gestures), joint attention, and pre-literacy skills (for example, listening comprehension)
Child health[a]	Height, weight, cortisol levels, body mass index, parental ratings of general health, and fine and gross motor skills
Children's long-term risk and economic well-being outcomes	
Long-term risk behaviors	Substance abuse, dropping out of high school, and teen pregnancy
Long-term economic well-being	Employment and home-ownership in adulthood
Interim outcomes	
Global child care quality	Scores on the Infant/Toddler Environment Rating Scale (ITERS) or Family Child Care Environment Rating Scales (FCCERS)
Structural features of care	Child-to-staff ratios; group size; caregiver qualifications; professional development; the physical environment and furnishings; schedules; personal care routines; and health, safety, and nutrition practices
Parent- or caregiver-child interaction	Sensitivity/responsiveness, learning and language supports/instruction and cognitive stimulation, positive regard/warmth, behavior guidance, support for peer interaction, and areas of concern in interactions
Parent or caregiver knowledge of child development	Caregiver's ability to identify developmental milestones
Global home environment	Home Observation for Measurement of the Environment (HOME) scores, language environment, cognitive stimulation, organization of the home, and safety

[a]Child health outcomes alone do not make a model eligible for inclusion in LITES, but child health outcomes are assessed for evidence of effectiveness.

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

APPENDIX B

DESCRIPTION OF DATABASES FOR LITES LITERATURE SEARCH

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

Appendix B. Description of databases for LITES literature search

Academic Search Premier. This multidisciplinary database provides full text for more than 4,500 journals, including full text for more than 3,700 peer-reviewed titles. Portable document format (PDF) files to 1975 or further are available for more than 100 journals and searchable cited references are provided for more than 1,000 titles.

Campbell Collaboration. The Campbell Collaboration website contains information about systematic reviews and randomized trials in education, social work and welfare, and criminal justice.

Child Care and Early Education Research Connections. The Child Care and Early Education Research Connections database is an online database where the early childhood community can share resources.

CINAHL with Full Text. CINAHL with Full Text is the world's most comprehensive source of full text for nursing and allied health journals, providing full text for nearly 600 journals indexed in *CINAHL*. This authoritative file contains full text for many of the most widely used journals in the *CINAHL* index with no embargo. Full-text coverage dates to 1981.

Cochrane Central Register of Controlled Trials. The Cochrane Central Register of Controlled Trials is a bibliography of controlled trials identified by contributors to the Cochrane Collaboration and others as part of an international effort to hand search the world's journals and create an unbiased source of data for systematic reviews.

Cochrane Database of Systematic Reviews. The Cochrane Database of Systematic Reviews contains full text articles and protocols focusing on the effects of health care. Data are drawn from evidence-based medicine and are often combined statistically (with meta-analysis) to increase the power of the findings of numerous studies that are each too small to produce reliable results.

Cochrane Methodology Register. The Cochrane Methodology Register (CMR) is a bibliography of publications that reports on methods used in the conduct of controlled trials. It includes journal articles, books, and conference proceedings; these articles are taken from the MEDLINE database and from hand searches. The database contains studies of methods used in reviews and more general methodological studies that could be relevant to anyone preparing systematic reviews. CMR records contain the title of the article, information on where it was published (bibliographic details), and in some cases a summary of the article. CMR is produced by the UK Cochrane Centre on behalf of the Cochrane Methodology Review Group.

Database of Abstracts of Reviews of Effects. The Database of Abstracts of Reviews of Effects (DARE) includes abstracts of published systematic reviews on the effects of health care from around the world, which have been critically analyzed according to rigorous criteria. This database provides access to quality reviews in subjects for which a Cochrane review might not yet exist.

E-Journals. E-Journals, makes use of the EBSCO*host* interface and offers a customized search environment.

ProQuest Dissertations and Theses. ProQuest Dissertations and Theses provides access to the world's most comprehensive collection of dissertations and theses, with more than 2.4 million dissertations and theses included from around the world. Each dissertation published since July 1980 includes a 350-word abstract written by the author. Master's theses published since 1988 include 150-word abstracts. Bibliographic citations are available for dissertations dating from 1637, and more than 65,000 new citations are added to the database every year.

EconLit. EconLit, the American Economic Association's electronic database, is the world's foremost source of references to economics literature. The database contains more than 785,000 records from 1969 to the present. EconLit covers virtually every area related to economics.

Education Research Complete. Education Research Complete is the definitive online resource for education research. Topics covered include all levels of education from early childhood to higher education and all educational specialties, such as multilingual education, health education, and testing. Education Research Complete provides indexing and abstracts for more than 1,840 journals, full text for more than 950 journals, and full text for more than 81 books and monographs and numerous education-related conference papers.

ERIC. Funded by the U.S. Department of Education (ED), ERIC is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All ED publications are included in its inventory.

MedLine. Medline is the United States National Library of Medicine's (NLM®) premier bibliographic database, providing information from the following fields: medicine, nursing, dentistry, veterinary medicine, allied health, and preclinical sciences. The MedLine database is the electronic counterpart of *Index Medicus®*, *Index to Dental Literature,* and the *International Nursing Index.*

PsycINFO. PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in the field of psychology. Journal coverage dates to the 1800s and includes international material selected from more than 1,700 periodicals. More than 60,000 records are added each year.

SocINDEX with Full Text. SocINDEX with Full Text is the world's most comprehensive and highest quality sociology research database. The database features more than 1,986,000 records with subject headings from a sociological thesaurus with more than 19,600 terms, designed by subject experts and expert lexicographers. SocINDEX with Full Text contains full text for 708 journals dating to 1908. This database also includes full text for more than 780 books and monographs and full text for 9,333 conference papers.

Scopus. Scopus is the world's largest abstract and citation database of peer-reviewed literature and quality web sources in the scientific, technical, medical, and social sciences. It covers more than 19,000 titles, articles in press, conference proceedings, and e-books.

SAGE Journals. This database provides access to the full text of articles in more than 500 leading journals published by SAGE, including all of the American Educational Research Association journals, as well as many leading titles in psychology, early childhood, and survey methodology.

APPENDIX C

CALL FOR STUDIES

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

LEARNING ABOUT INFANT AND TODDLER EARLY EDUCATION SERVICES (LITES): IDENTIFYING WHAT WORKS AND ADVANCING MODEL DEVELOPMENT

2014 CALL FOR STUDIES

SUBMISSION DEADLINE: MARCH 25, 2014

Mathematica Policy Research® seeks studies for a review that will assess the evidence base of out-of-home early care and education (ECE) models for infants and toddlers (from birth to age 3). The review is being conducted by Mathematica for the Office of the Assistant Secretary for Planning and Evaluation (ASPE), in partnership with the Office of Planning, Research & Evaluation (OPRE) within the U.S. Department of Health and Human Services. It will be used to help inform policy, new initiatives, and program directions at the federal level. Submissions are due by March 25, 2014.

Background

A growing body of research indicates that high quality early learning experiences can promote young children's development and help to reduce achievement gaps. However, little is known about what works for children from birth to age 3 in terms of early learning services designed to support children's cognitive, language, and social-emotional/behavioral development. Increasingly, federal policymakers are using research evidence to inform decision making and funding evidence-based program models as part of efforts to make smarter investments in education, health care, and social services (Burwell et al., 2013; Haskins & Baron, 2011). To help identify effective and replicable program models of out-of-home ECE services for infants and toddlers, ASPE, in partnership with OPRE, is conducting a systematic review of the evidence base.

Purpose

The purpose of this review is to identify replicable program models that have demonstrated evidence of effectiveness for supporting infant and toddler early learning in the domains of cognition, language, and/or social-emotional/behavioral development in out-of-home ECE settings. These settings may include ECE centers and family child care homes. The review will include program models that provide (1) direct early learning services to infants and toddlers in out-of-home ECE settings and/or (2) indirect early learning services through professional development for adult out-of-home caregivers designed to support infant and toddler early learning.[31] Caregivers may include teachers/caregivers in infant and toddler ECE classrooms and

[31] Program models that provide infrequent or supplemental home visits may be considered for inclusion in the review, but the primary service setting must be out-of-home care. Program models that provide supplemental services in areas such as nutrition, health and developmental screening, supports for parents, and referrals to other community resources may be considered for inclusion in the review, but the primary focus of services must be supporting infant and toddler early learning delivered outside the child's home. Indirect services—such as parenting, family self-sufficiency, or referral services—will not be included because they do not target children's early learning in out-of-home care settings.

family child care providers. To be considered replicable, program models must at a minimum provide a defined set of infant and toddler early learning service components or professional development services to help caregivers support infant and toddler early learning.

For the purpose of this review, infants and toddlers are children from birth to age 36 months.[32] Services must be targeted broadly to infants and toddlers and/or their adult out-of-home caregivers. Program models targeted narrowly to infants and toddlers with diagnosed disabilities or specific medical conditions will not be included in the review.[33] However, services targeted to broad groups of at-risk infants and toddlers (for example, children from low-income families or low-birth-weight children) will be eligible for inclusion.

Eligibility

This call for studies aims to identify unpublished manuscripts (past or recent), conference papers, new publications (currently in press), or manuscripts with new analyses of already published work that are not included in existing research databases. Apart from the call for studies, the Mathematica team will conduct keyword searches of electronic databases and other search activities. The review will include all relevant studies from these searches, supplemented with additional studies identified through this call.

Studies submitted in response to this call should:

- Focus on program models that provide direct early learning services to infants and/or toddlers in out-of-home care or indirect early learning services through professional development for out-of-home caregivers. (Professional development services must involve intervening directly with caregivers and take place in the caregiving setting or a similar setting.)
- Include study samples in which at least half of the children were initially enrolled in services at age 30 months or younger.
- Have been prepared or published in 1980 or later.
- Provide the name and a detailed description of the program or model being evaluated, as well as the study design, analysis methods, and findings. Slide presentations and abstracts alone should not be submitted as they will not provide sufficient detail for the review.
- Target at least one child outcome in at least one of the following domains: cognitive development, social-emotional/behavioral development, and/or language development.
- Be accessible to the public through a website, as a published article or book chapter, or upon request from the study author. (Mathematica will not publically distribute studies; however,

[32] Programs that enroll families before the child's birth may be included in the review, as long as the primary focus of services is supporting children's early learning in out-of-home ECE settings.

[33] The Administration currently has substantial investments in special education and supporting the development of children with disabilities. The focus of this review is to identify effective program models for supporting early learning for a broad range of infants and toddlers. However, if subgroup impacts are reported for children with diagnosed disabilities or specific medical conditions in a study of an intervention that targets infants and toddlers broadly, subgroup impacts will be reported in the review.

to ensure transparency of the review, the manuscript should be available upon request and should not be confidential.)

Submission Instructions

Submissions should include the following:

- An electronic version of the study in MS Word, PDF, or RTF format
- A cover email noting:
  - Contact information for the lead or corresponding author
  - The name of the out-of-home ECE program for infants and toddlers being evaluated
  - The study design—randomized controlled trial, matched comparison group design, regression discontinuity design, single case design, nonexperimental design (such as pre-post or correlational), or implementation study

Submissions should be emailed to LITES@mathematica-mpr.com

The deadline for submissions is March 25, 2014.

Submitters will receive acknowledgment of receipt of their submission but no indication of the possible inclusion of their study in the review.

References

Burwell, Sylvia M., Cecilia Munoz, John Holdren, and Alan Krueger. "Next Steps in Evidence and Innovation Agenda." Memorandum to the Heads of Departments and Agencies, July 26, 2013. Available at http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-17.pdf.

Haskins, Ron, and Jon Baron. "Building the Connection Between Policy and Evidence: The Obama Evidence-Based Initiatives." London, UK: NESTA, September 2011.

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

APPENDIX D

SCREENING DISPOSITION CODES

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

Table D.1. Screening disposition codes

Exclude publications that are not studies • Not a study • Handbook or conference proceedings • Supplemental materials
Non-English study • Study is not written in English
International and not policy relevant • Program is delivered in a developing-world context
Publication date is out of range • Publication date is ineligible
Not a primary study • This citation is not a primary study
Studies on services not relevant to the review • Early learning or development is not a substantial goal of the program • Program targeted to children with diagnosed disabilities or specific medical conditions
ECE target population out of range • Study sample does not include any children enrolled before age 36 months • Study sample does not include at least 50 percent children enrolled before age 30 months (this status will be applied by reviewers rather than screeners)
Not possible to attribute effects solely to the model of interest • Not possible to attribute effects solely to the model of interest
Study does not examine a replicable program • Did not include a defined package of replicable program components
Subgroups out of scope • Study only reported on subgroups that were not the LITES pre-identified subgroups of interest
No eligible outcomes • No eligible child outcomes
Design screening • Ineligible study design
Additional dispositions • Study passes screens (will note in SharePoint if direct early learning services or professional development) • Hold—unclear whether out-of-home ECE early learning services are involved, otherwise passes • Hold—out-of-home ECE early learning services are involved, unclear which program model, otherwise passes • Hold—need team management review • Hold—need full text • Study is not the most recent and complete version available • Could not obtain the text of this study

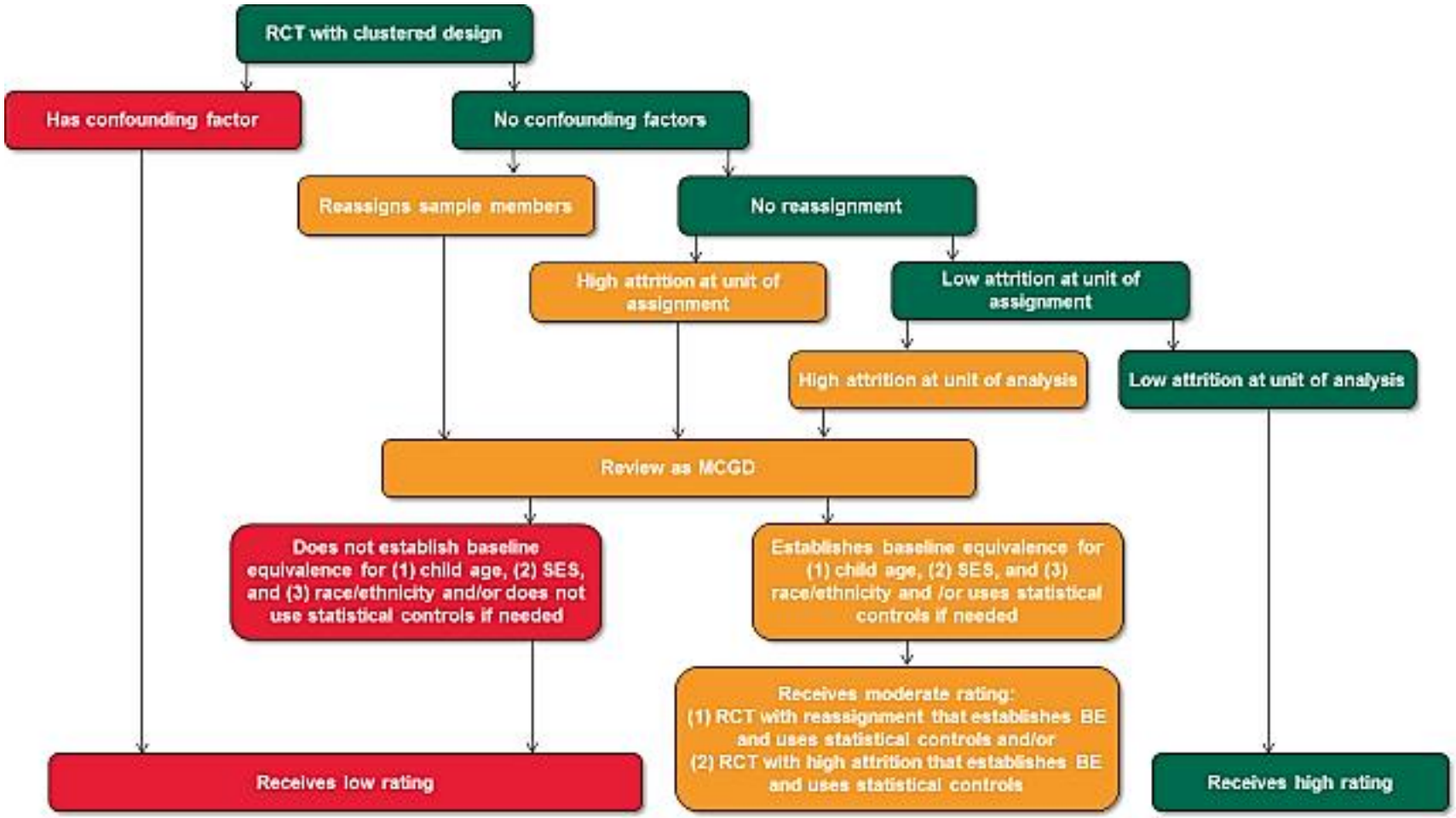
This page has been left blank for double-sided copying.

APPENDIX E

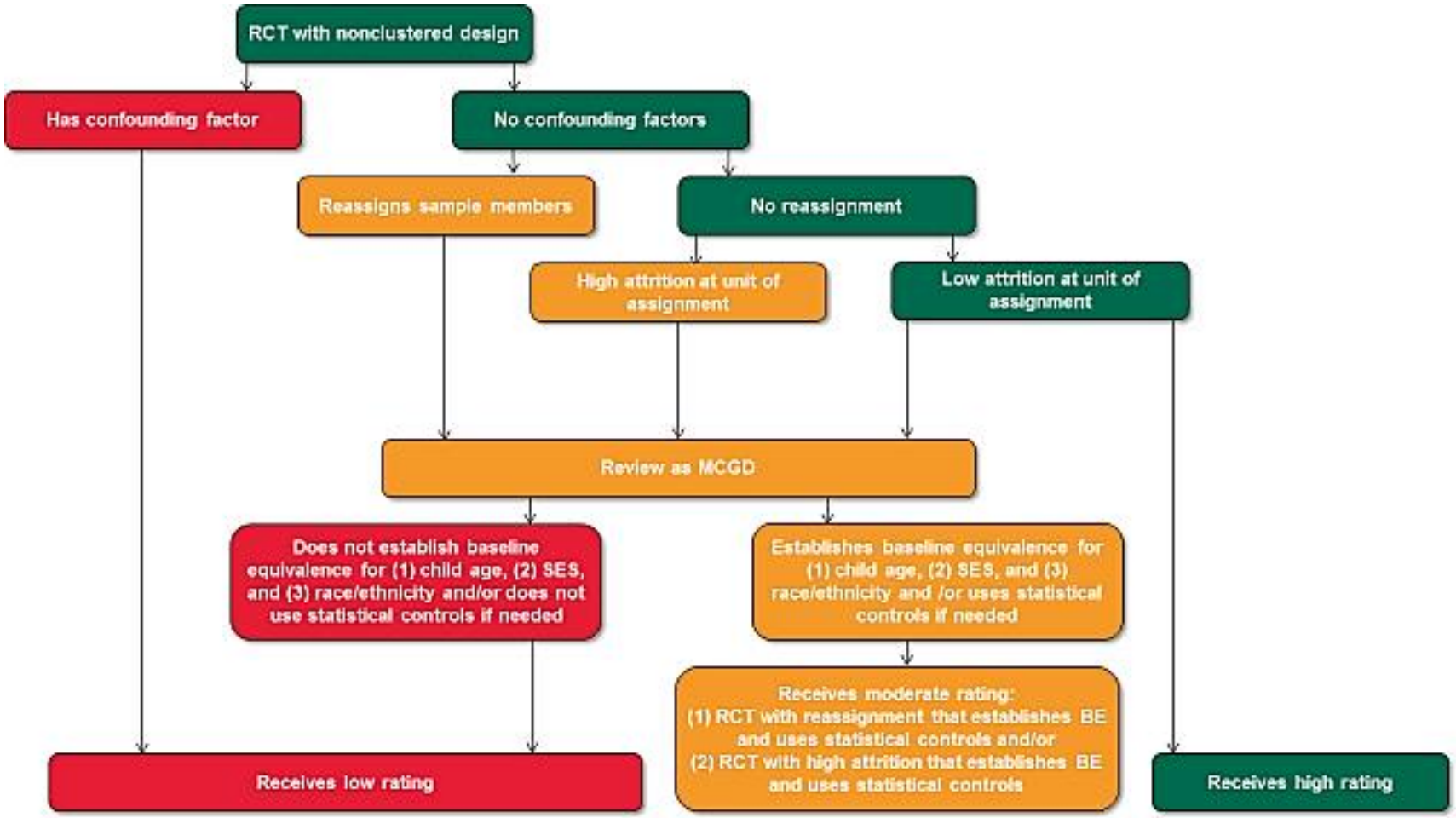
STUDY QUALITY RATING DECISION TREES

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

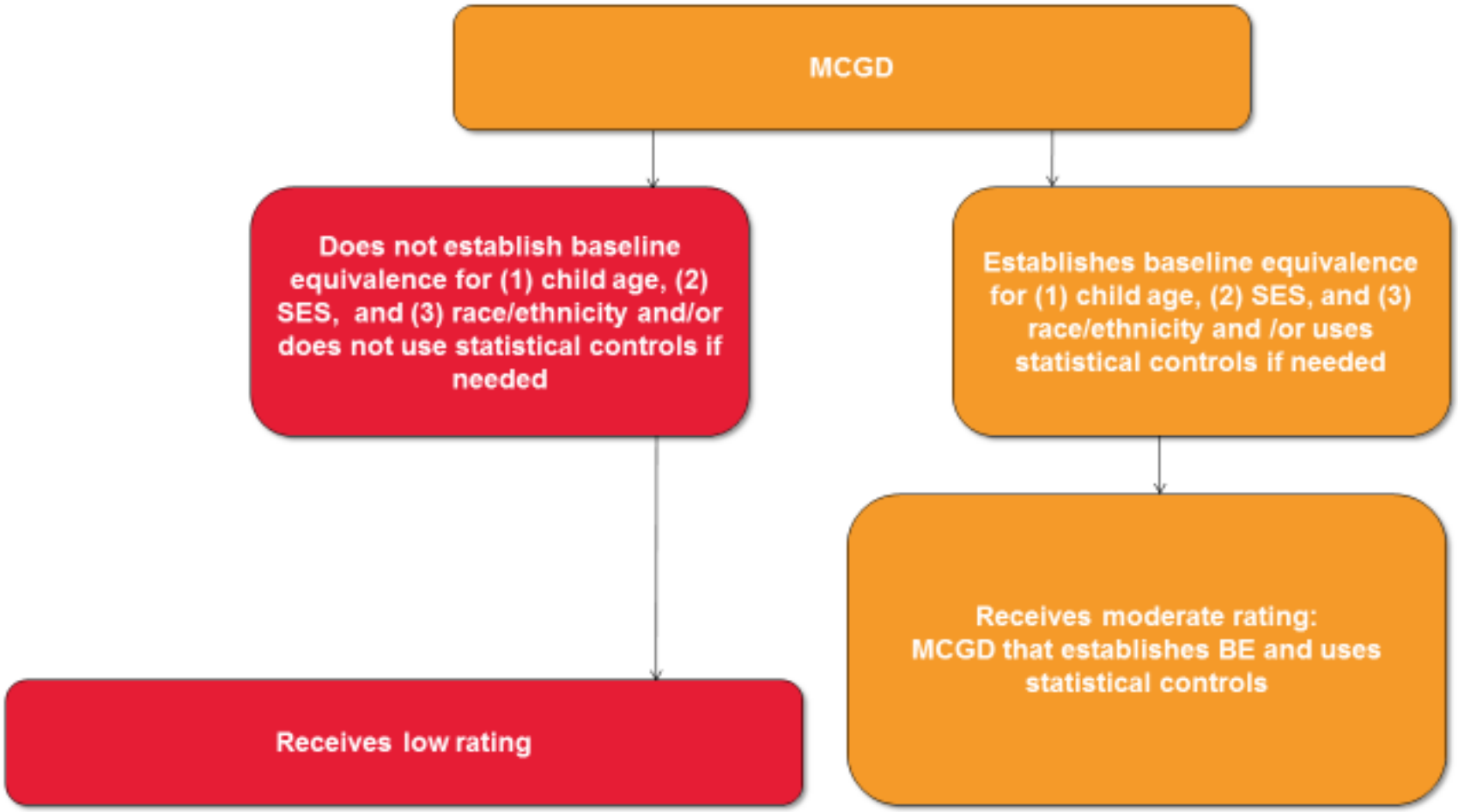
Randomized controlled trial (RCT) with clustered design decision tree



Randomized controlled trial (RCT) with nonclustered design decision tree



Matched comparison group design (MCGD) decision tree



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

APPENDIX F

LITES REVIEWER TRAINING AGENDA

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

APPENDIX F. LITES REVIEWER TRAINING AGENDA

I. Introduction

1. Project team roles and responsibilities
2. Description of LITES systematic review

II. Review process

1. Two sequential reviews
   a. R1's role: Complete Study Review Guide (SRG) and SharePoint record
   b. R2's role: Thoroughly check R1's SRG and rating, document any disagreements, come to an agreement with R1 or discuss with project leadership, create master SRG after reconciler confirms final rating
   c. Reconciler: Confirm final rating
   d. Team leadership: Quality assurance check
2. Reviewer assignments
3. Review time frame
4. Author queries

III. The LITES criteria to assess study quality and evidence of effectiveness

1. Study quality criteria
   a. Eligible study designs
   b. Study ratings
   c. Confounding factors
   d. Reassignment
   e. Attrition
   f. Baseline equivalence
   g. Statistical controls
   h. Multiple ratings
2. Evidence of effectiveness

IV. Documenting the review

1. Use of SharePoint
   a. Tracking review progress
   b. Basic information recorded from study reviews
   c. SharePoint fields
2. The SRG
   a. Description of each field in the SRG, with detailed instructions on how to complete.
3. Overview of other review matrices
   a. Master outcomes matrix
   b. Program model matrix: used to assign an evidence of effectiveness rating
      (1) Information recorded that spans multiple studies (for example, sustained/replicated effects)
   c. Program model summary

V. Contents of training manual

1. Agenda
2. The review protocol
3. Matrices

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

APPENDIX G

LITES AUTHOR QUERY TEMPLATE

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

APPENDIX G. LITES AUTHOR QUERY TEMPLATE

Dear [Recipient]:

The Office of the Assistant Secretary for Planning and Evaluation within the U.S. Department of Health and Human Services has contracted with Mathematica Policy Research to conduct *Learning about Infant and Toddler Early Education Services* (LITES), a systematic review to identify effective models of early care and education (ECE) services for infants and toddlers.

The purpose of this letter is to notify you that as part of our review of studies of ECE services, we are reviewing the following study for possible inclusion in LITES reports and deliverables: [Insert study citation]

Would you please help us better understand the data in your study by responding to the following requests?

A. Please provide the following information: [Insert Questions]

[DELETE IF A TABLE DOES NOT ACCOMPANY THE QUERY]

B. Please complete the attached table/tables about your study. The table/tables asks/ask for information on measures and outcomes of interest for this LITES study review.

If possible, could you please provide this information by [Date – two weeks from date of letter]? We recognize this is not much time, but we are trying to maintain a brisk pace for our project. If we do not receive a response from you by [Date], we will proceed with the information we have. Please mail or fax your responses to Lauren Murphy, Research Analyst, Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393, Fax: (609) 799-0005. If you prefer, you can email your responses to LITESaq@mathematica-mpr.com.

[IF THE STUDY IS NOT PUBLICLY AVAILABLE] If your study is not publicly available, we ask that LITES may have a non-exclusive, royalty-free license to use the study for Federal purposes associated with LITES. This license includes other information and correspondence submitted by the author for the LITES process. LITES agrees that no other use of the study or reports and information will be made without prior permission. In addition, LITES is not responsible for responding to requests by third parties for a copy of the author's study if the study is not publicly available. LITES will forward such requests to the author, as appropriate.

Please do not hesitate to contact us if you have any questions about the query. Thank you very much for your help.

Sincerely,

Diane Paulsell

Project Director, LITES

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

APPENDIX H

RESEARCH TERMS GLOSSARY

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

APPENDIX H. RESEARCH TERMS GLOSSARY

A

Absolute value. The value of a number, as a distance from zero, disregarding whether the number is positive or negative. For example, the absolute value for both +4 and -4 is 4.

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

B

Baseline. The study's onset.

Baseline equivalence. Occurs when the intervention and comparison groups have similar characteristics (such as race and age) at the study's onset. For LITES, baseline equivalence was established when no statistically significant differences were detected on required measures at baseline.

C

Clustered randomized controlled trial (clustered RCT). Clusters (such as child care centers) are randomly assigned to the intervention.

Comparison group. A group with characteristics similar to those of intervention group members, except that they 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 a comparison group is to the intervention group, the more likely it is that any difference in outcomes between the two groups can be attributed to the intervention.

Confounding factor. Occurs when an aspect of the study design, other than the model of interest, aligns with the intervention or comparison group, making it impossible to measure unbiased impact. For example, if one classroom caregiver administers all program ECE services, it is impossible to distinguish the effectiveness of that person from the effectiveness of the program. Confounding factors may also arise from systematic differences in the way data are collected from participants in the intervention group versus the comparison group. For example, participants may report information differently to someone they know than to someone they do not know. Familiarity with the data collector may change the way participants answer the questions. The presence of confounding factors can impede the ability of a study to capture an estimate of the actual effect of a program (that is, an unbiased impact).

Cronbach's coefficient alpha. An estimate of internal consistency reliability that indicates how well groups of items in an assessment "hang together" and contribute to measurement of the same construct. The estimate captures the extent to which the separate items on the measure all seem to move in the same direction (that is, if a person is high on one item of a construct, they rate themselves high on all of the items related to that construct on a measure). The greater the similarity among items, the higher the reliability (and thus the higher the value of Cronbach's coefficient alpha). Values of the alpha can range from -1.0 to 1.0, with greater values indicating stronger internal consistency.

D

Differential attrition. Differential attrition rate is the absolute value of the difference between the attrition rates in the intervention and comparison groups.

E

Effect size. A measure of the magnitude of the difference between the intervention group and the comparison group. The effect size shows the magnitude 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 using different units. In the LITES review, a negative value indicated 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 indicated that the outcomes for the intervention group were greater than those for the comparison group. Values of 0 (referred to as a neutral effect) indicated there was no difference, on average, between the intervention and comparison groups.

F

Favorable effect. An estimated impact on an outcome measure in a direction that is beneficial for children and parents. This impact could be positive or negative, and is determined to be "favorable" based on the end result. For example, a favorable impact could be an increase in children's vocabulary or a reduction in harsh parenting practices.

Follow-up. A time point after the onset of the intervention for measuring participant outcomes.

I

Internal validity. A study's ability to isolate the effects of an intervention from other factors that may influence participants' outcomes.

Intervention group. The sample members who receive the early care and education services or program of interest.

M

Matched comparison group design (MCGD). A study design in which sample members (children, parents, or families) are selected for the intervention and comparison conditions in a nonrandom way.

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

N

Null effect. An effect that is neither a significant or substantial favorable effect nor a significant or substantial unfavorable effect.

O

Outcome domain. A group of related outcomes that measure the same or similar constructs. The LITES review includes three primary child outcome domains: (1) cognitive development, (2) social-emotional/behavioral development, or (3) language development. Child health outcomes such as height, weight, gross and fine motor skills, and hospitalizations were reported if present in a study of a model that had at least one study with child outcomes in a cognitive, social-emotional/behavioral, or language domain. The LITES review also included long-term risk and economic well-being outcomes and several interim domains.

Overalignment. When outcome measures more closely align to one of the study groups than the other and could bias a study's results.

Overall attrition. The total number of sample members who are not participating at follow-up.

P

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

R

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

Reassignment. Compromising or violating random assignment—for example, children being switched from the comparison group to the intervention group after random assignment. If these

children's outcome data were included as part of the intervention group's results, the study would suffer from reassignment and could not be reviewed as an RCT.

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

Replicated effect. An effect that is statistically significant ($p \leq 0.05$) or has an effect size greater than or equal to 0.2 standard deviations and is measured in two or more non-overlapping analytic study samples.

S

Sample. Persons (children, caregivers, or families) included in the study. For the LITES review, sites that were analyzed separately were considered separate samples.

Significant effect. An impact estimate that is statistically significant with $p \leq 0.05$.

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

Standard deviation. A measure of the spread or variation of values in the sample. The standard deviation approximates the distribution around the mean with 68 percent of the sample having values that are between one standard deviation below the mean and one standard deviation above the mean. Smaller standard deviations indicate that the values for individual sample members are closer to the mean, whereas larger standard deviations indicate there is more variation in values.

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

Statistical controls. Methods of adjusting for characteristics that may differ between the intervention and comparison groups at baseline to make the groups more comparable.

Statistical significance. An indication of the probability that the observed finding was obtained by chance (when there is not a real relationship in the population). If the *p*-value is equal to or less than a predetermined cutoff (in the LITES review, 0.05), the finding is considered statistically significant because it has a low probability of having occurred by chance (5 percent or less).

Substantial effect. An impact estimate that has an effect size greater than or equal to 0.2 standard deviations in absolute value.

Sustained or delayed effect. An effect that is statistically significant ($p \leq 0.05$) or has an effect size greater than or equal to 0.2 standard deviations and is measured one year or more after the end of the intervention.

www.mathematica-mpr.com

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

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research