Solving the Early Childhood Program Evaluation Paradox: Re-envisioning Evidence via the CIRCLE Framework

Philip Fisher & Kathryn Beauchamp

The Stanford Center on Early Childhood
Working Paper Series

September 2025



Abstract

Despite decades of investment in early childhood programs and an evidence base demonstrating their positive impacts, the field continues to face significant challenges in the selection and implementation of evaluation methodology. This creates issues for policy makers, advocates, and funders who seek to understand the impact of early childhood investments. It also results in impediments in identifying the optimal approaches to supporting young children, their parents, and other adults in their lives and understanding variations in impacts of investments. This paper examines contemporary issues in early childhood program evaluation, focusing particularly on the limitations of randomized controlled trials, overreliance on average treatment effects, and insufficient attention to variation in program impacts across and within different populations. Drawing on recent advances in developmental science and evaluation methodology, we propose the Continuous Improvement Rapid Cycle Learning and Evaluation (CIRCLE) framework that embraces a continuum of evaluation approaches, prioritizes understanding of impact variation, and emphasizes greater precision in theories of change. The paper concludes with recommendations to advance the field of early childhood program evaluation toward more precise, more equitable, and ultimately more impactful interventions.

Keywords: early childhood intervention, program evaluation, randomized controlled trials, differential impacts, developmental science

---

Author note: This working paper was prepared by the authors with the assistance of claude.ai. Specifically, we used claude.ai to summarize prior work previously published by one or more of the current authors on the state of early childhood evaluation research, and to condense content in the CIRCLE Framework manuals and related documentation.



Solving the Early Childhood Program Evaluation Paradox: Re-envisioning Evidence via the CIRCLE Framework

More than half a century of research on early childhood programs, policy, and practice has produced substantial evidence demonstrating the value of investing in the early years. From pioneering programs like the Perry Preschool Project and the Carolina Abecedarian Project to national initiatives like Head Start, research has shown that high-quality early childhood programs can improve outcomes across cognitive, social-emotional, behavioral, and health domains (Shonkoff & Fisher, 2013). The economic case for such investments is similarly compelling, with cost-benefit analyses showing significant returns (Heckman et al., 2010; Belfield et al., 2006; Hahn & Barnett, 2023).

Despite these achievements, the field faces a paradox. While the evidence base supporting early intervention has expanded dramatically, population-level impacts remain elusive, and meta-analyses suggest that effect sizes across many programs remain modest (Duncan & Magnuson, 2013; Whitaker et al., 2023). This disconnect between the promise demonstrated in well-controlled studies and the challenges of achieving broad impact raises important questions about how we evaluate early childhood programs, make meaning of empirical evidence from evaluation research, and translate findings into practice (Schorr & Farrow, 2011).

The purpose of this paper is to examine contemporary challenges in evaluating early childhood programs and to propose innovative approaches for advancing the field. We focus on three interrelated issues: (1) limitations in the dominant evaluation paradigm that relies heavily on randomized controlled trials (RCTs); (2) overemphasis on average treatment effects at the expense of understanding variations in impact; and (3) insufficient precision in program theories and measurement. For each challenge, we identify underlying concerns and propose alternative approaches that build on recent advances in developmental science and evaluation methodology.

Our central argument is that moving beyond the current state of early childhood program evaluation requires a shift away from the dichotomous question of whether programs "work" toward a more nuanced understanding of for whom and under what conditions specific intervention components are most effective.

We view evaluation as a tool for continuous learning and program improvement rather than for delivering conclusive proof that a particular program or strategy works. This shift has important implications for program design and evaluation, for guiding funding decisions, and for translating research into practice and policy.

The Historical Context: Evolution of Early Childhood Intervention Research

The field of early childhood program evaluation research emerged in the United States in the 1960s, during the War on Poverty era, and was shaped by a convergence of developmental theory and social policy (Fisher et al., 2020). The justification for funding early childhood programs was grounded in an emerging body of research demonstrating the malleability of development and the impact of early experiences on long-term outcomes.

The creation of Head Start in 1965 was an unprecedented moment in the US, in which the federal government committed to a national early childhood strategy of reducing educational and economic opportunity gaps between wealthy and low-income individuals. Over the ensuing decades, as Head Start expanded and evolved to include Early Head Start, alongside the development of other early childhood programs, evaluation research methodologies in the field also advanced. The Perry Preschool Project, launched in 1962, and the Carolina Abecedarian Project, initiated in the 1970s, established the foundational evidence for early childhood programs via rigorous designs and longitudinal follow-ups showing sustained benefits into adulthood (Schweinhart et al., 2005; Campbell et al., 2012).

Evaluation practices evolved from primarily descriptive studies to the gold standard of randomized controlled trial designs (RCTs) that sought to establish causal links between interventions and outcomes.

By the early 2000s, RCTs had become the predominant approach to establishing evidence of program effectiveness (Raudenbush, 2018). This shift was part of a broader movement toward evidence-based practice across social policy domains. The emphasis on RCTs was reinforced by the creation of organizations like the What Works Clearinghouse, which privileged evidence from randomized trials in determining which programs were considered "evidence-based" (What Works Clearinghouse, 2020).

Parallel to these developments in evaluation methodology, the science of early childhood development was also advancing rapidly. Research in neuroscience, molecular biology, genomics, and epigenetics provided increasingly sophisticated insights into how early experiences shape developmental trajectories (Shonkoff & Fisher, 2013). These scientific advances offered new opportunities to understand the underlying mechanisms through which early interventions influence development. However, as Horn et al. (2020) observe, the translation of this expanding knowledge base into innovative intervention approaches has been limited.

Despite increasingly sophisticated understanding of developmental processes, the predominant evaluation paradigm continues to focus on whether programs work on average, rather than unpacking the mechanisms through which they work or examining variations in impact across different populations.

Limitations of the Status Quo in Early Childhood Program Evaluation

In the sections that follow, we describe three challenges in the field of early childhood program evaluation research. First, we describe the limitations of reliance on RCTs as the primary source of evidence of impact. Second, we consider the implications of measuring average program impacts without considering variations in impact among those receiving programs. Third, we discuss the lack of specificity and precision in much of the early childhood program evaluation research. We then propose a novel framework that has the potential to increase the impact of early childhood programs at the population level and also provide policy makers, advocates, funders, and other stakeholders with tools to more readily determine the return on early childhood investments.

Limitations of the RCT Paradigm

The RCT has long been considered the gold standard for establishing causal evidence in early childhood intervention research. The logic is straightforward: by randomly assigning participants to treatment and control conditions, researchers can attribute observed differences in outcomes to the intervention with a high degree of confidence. This approach has been instrumental in building the evidence base for early childhood programs and informing policy decisions.

However, as Fisher et al. (2020) point out, the discourse in this area "too often endorses RCTs as the only scientifically viable approach to establishing evidence" (p. 277). This reliance on RCTs as the sole arbiter of program effectiveness has at least three limitations in the context of early childhood intervention.

First, RCTs suffer from many of the same threats to internal and external validity as other research designs (Grossman & Mackenzie, 2005). Despite random assignment, baseline differences between groups can emerge, particularly in smaller samples. Post-randomization experiences can also differ between groups in ways that introduce confounding variables. These issues are particularly salient in early childhood settings, where it is often impossible to blind participants or providers to group assignment (Cook, 2018).

Second, the practical challenges of implementing RCTs in community settings are considerable. RCTs require substantial resources, time, and control over implementation conditions that may



not be feasible in many real-world contexts. This creates a bias toward evaluating programs that are amenable to randomization and backed by substantial evaluation funding, rather than those that might be most promising or innovative. It also may raise equity issues, because community-based programs that are under-resourced and/or are located in low-income settings may lack access to evaluation researchers and other resources needed to conduct an RCT.

Third, and perhaps most important, the emphasis on RCTs often leads to an all-or-none approach to evidence, where programs are categorized as either "evidence-based" or not based on the presence of statistically significant average effects on outcome variables of interest (whether or not such outcomes were specified a priori by the researchers). This dichotomous framing obscures the complexity of program impacts and limits attention to understanding the conditions under which programs are most effective.

As Deaton and Cartwright (2018) argue, "RCTs, like other methods, have their uses, but the claims that are made for them are sometimes exaggerated, and the automatic preference for RCTs over other methods is not justified" (p. 2). We argue that what is needed is not a rejection of RCTs but a more nuanced approach that views them as one tool within a broader toolkit of evaluation methods.

Fisher et al. (2020) propose a continuous quality improvement approach to evaluation that recognizes a continuum of evidence. In this approach, evaluation designs are matched to the stage of program development and the specific questions being addressed. Early-stage programs might appropriately be evaluated using less rigorous designs focused on feasibility and implementation, while more established programs might warrant RCTs to test specific causal hypotheses.

This approach aligns with calls for greater methodological pluralism in social intervention research (Pawson & Tilley, 1997; Ludwig et al., 2011). It recognizes that different evaluation questions require different methodological approaches and that understanding the complex processes through which interventions influence development requires multiple forms of evidence. Moreover, other scholars have proposed that this sort of preliminary evidence can provide tools for optimization of interventions before investing in large-scale RCTs—thereby potentially increasing program impacts and efficiency and driving down program costs in the long run.

Overemphasis on Average Treatment Effects

The second major challenge in contemporary early childhood program evaluation is the overemphasis on average treatment effects (ATEs) as the primary metric of program impact. In conventional approaches to program evaluation, the focus is on comparing mean outcomes in the treatment group to mean outcomes in the control group, with statistical significance as the key criterion for success. This focus on ATEs obscures important variations in program impact across different subgroups and contexts. As Fisher et al. (2020) note, "examining variations in



impact is at least as important as examining average effects" (p. 284). Understanding these variations is critical for several reasons.

First, even when programs show statistically significant average effects, variations in impact always occur. Some participants may benefit substantially while others show little or no improvement. For example, the Head Start Impact Study found considerable variation in program effects across domains and across different subgroups of children (Puma et al., 2012). Understanding these patterns of differential impact can provide insights into how programs might be adapted to better serve those who benefit least.

Second, the focus on statistical significance rather than effect size magnitude can lead to misleading conclusions about program effectiveness. As Duncan and Magnuson (2013) observe, most early intervention effect sizes fall below 0.5, which corresponds to a relatively small percentage of individuals showing meaningful improvements relative to what would be expected without intervention. This disconnect between statistical significance and practical significance is rarely acknowledged in discussions of program effectiveness.

Third, the assumption that average effects can be extrapolated to individuals or to different populations ignores the importance of contextual and individual factors in shaping program impacts. Cultural differences, community characteristics, and individual needs all influence how programs are experienced and their ultimate effects.

The limitations of the emphasis on ATEs point to the need for more attention to moderators of intervention impacts—that is, the identification of characteristics of those receiving the program (e.g., baseline risk factors), characteristics of those delivering the program (e.g., training or cultural competence), or features of the implementation context (e.g., community resources) that are reliably associated with variations in observed outcomes.

Fisher et al. (2020) argue for a priori specification of hypothesized moderators that can be measured and analyzed in the context of program evaluations.

Recent advances in statistical methods offer promising approaches for examining impact heterogeneity. Techniques such as causal forests, Bayesian additive regression trees, and other machine learning approaches allow researchers to identify patterns of differential treatment effects without imposing strong a priori assumptions about where these differences might lie (Athey & Imbens, 2016). These methods complement more traditional approaches to moderation analysis and offer opportunities for discovering unexpected patterns of program impact.

Moving beyond ATEs also requires greater attention to individual-level change. Statistical methods such as the reliable change index (Jacobson & Truax, 1992) provide ways to assess whether observed changes for individuals are statistically reliable and clinically meaningful. As

Estrada et al. (2019) demonstrate, even a moderate effect size at the group level may correspond to reliable change for only a small proportion of individuals. Understanding these individual patterns of response can provide insights that are obscured in analyses focused solely on group averages.

Lack of Specificity and Precision in Theory and Measurement

The third challenge in contemporary early childhood program evaluation relates to the lack of specificity and precision in program theories and measurement. Early childhood programs often suffer from what has been called the "black box" problem—a lack of clarity about how program activities are hypothesized to be causally linked to outcomes (Correnti & Rowan, 2007; Harachi et al., 1999).

As Schindler et al. (2019) found in their meta-analysis of early childhood program evaluations, a disconnect often occurs between stated program goals and the outcomes that are actually measured (see Figure 1). Many programs focus on downstream outcomes without adequately measuring or addressing the proximal mechanisms through which these outcomes are expected to be achieved.

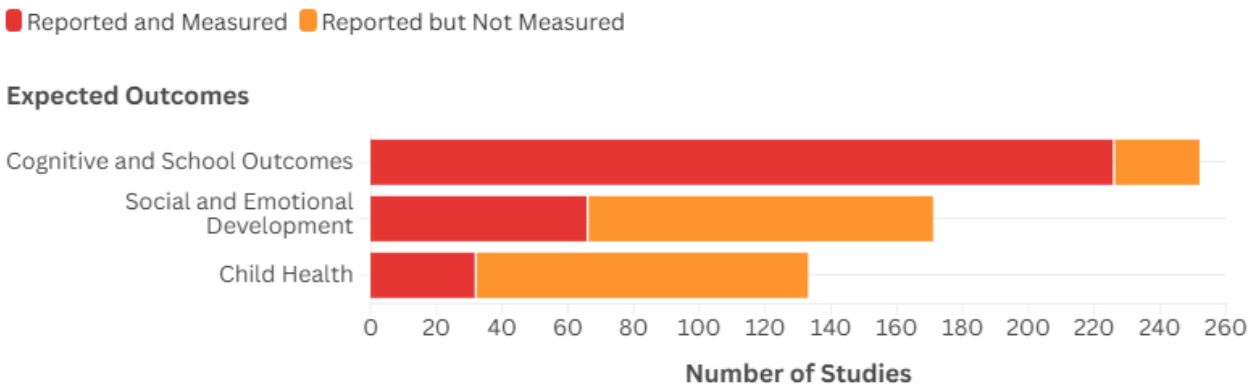


Figure 1. Illustrates the disconnect between reported outcomes and those that are actually measured in early childhood program evaluations. Adapted from Schindler et al. 2019.

Moreover, the conceptual models guiding program design and implementation are often insufficiently specified to guide precise measurement or to inform adaptations when expected impacts are not achieved. Theories of change or logic models are frequently treated as compliance requirements rather than as tools for guiding program development and evaluation.

In addition, measurement in early childhood program evaluation often lacks the precision needed to detect meaningful change. The field has relied heavily on broad measures of developmental domains rather than targeted assessments of specific skills or behaviors that programs aim to influence. This approach limits our ability to understand the specific pathways through which programs affect development. Relatedly, impacts are often assessed via teacher or parent reports, which may be affected by various biases, rather than direct assessment of

child outcomes. Additionally, tools designed primarily for screening are sometimes employed as outcome measures.

Addressing these challenges requires greater specificity and precision in both theory and measurement. Previously, we proposed a parsimonious template for theories of impact that distinguishes among program strategies (actions taken by the program), targets (domains directly acted upon by the program), outcomes (domains not directly targeted but expected to be influenced), and moderators (Fisher et al., 2020). This approach encourages program developers and evaluators to be explicit about the causal chains linking program activities to expected outcomes.

It is important to recognize that greater precision in measurement requires careful alignment between program targets and assessment tools. This alignment should be guided by advances in developmental science that provide increasingly sophisticated understanding of the specific skills and capacities that underlie broader developmental domains. For example, rather than simply measuring "executive function" as a global construct, programs might target and assess specific components such as working memory, inhibitory control, or cognitive flexibility.

The use of shared, standardized measures across studies can also enhance precision by facilitating comparison across programs and contexts. Efforts like the National Institutes of Health Toolbox (Weintraub et al., 2013) provide normed measures of cognitive, emotional, motor, and sensory functions that can be used across diverse studies. However, these should be complemented by measures that are sensitive to the specific targets of individual programs.

Finally, greater precision in understanding program impacts requires attention to implementation quality and fidelity. Variations in how programs are delivered can substantially influence their effectiveness. Systematic approaches to measuring implementation, such as those proposed by Durlak and DuPre (2008), can help to clarify the conditions under which programs are most effective.

Toward a New Framework for Early Childhood Program Evaluation

Building on the issues identified above, we propose a framework for early childhood program evaluation that embraces methodological pluralism, prioritizes understanding of impact variation, and emphasizes greater precision in theory and measurement. This framework, which we refer to as Continuous Improvement Rapid Cycle Learning & Evaluation (CIRCLE; see Figure 2), consists of three core components.

First, the CIRCLE Framework adopts a continuous quality improvement perspective on evaluation. Rather than viewing evaluation as a one-time, high-stakes activity to determine whether a program "works," this perspective conceives of evaluation as an ongoing process of learning and refinement. Different evaluation questions and methods are appropriate at different stages of program development, and the rigor of evaluation should be calibrated to the state of knowledge and the specific questions being addressed. For example, in the early stages of program development, exploratory and qualitative methods focused on feasibility,

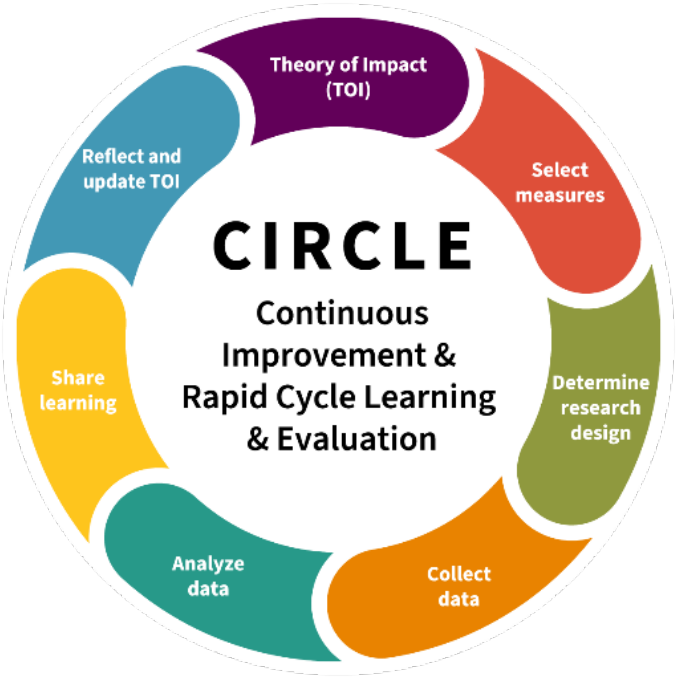


Figure 2. The CIRCLE Framework

acceptability, and initial evidence of impact may be most appropriate. As programs become more established, increasingly rigorous methods can be employed to test specific causal hypotheses. This staged approach allows for rapid cycles of innovation and adaptation rather than lengthy delays between program development and definitive evaluation.

Second, the CIRCLE Framework prioritizes understanding of variations in impact. This requires moving beyond the question of whether programs work on average to examine for whom and under what conditions they are most effective. Key elements of this approach involve: (a) specification of hypothesized moderators based on developmental theory and prior research; (b) use of appropriate statistical methods to examine patterns of differential impact; (c) attention to both group-level and individual-level patterns of response; and (d) translation of findings about impact variation into targeted adaptations to enhance program effectiveness for subgroups that show limited benefit.

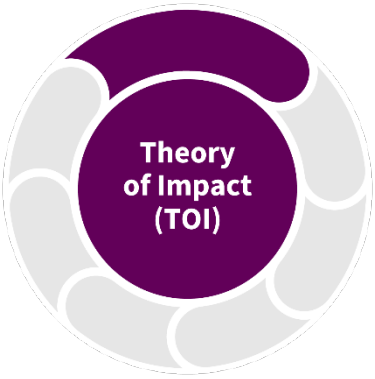
Third, the CIRCLE Framework emphasizes greater precision in theory and measurement. This begins with the development of a detailed theory of impact that specifies the causal chains linking program activities to proximal targets and distal outcomes. It requires that measurement strategies are aligned with the program theory of impact. It emphasizes the use of multiple methods (quantitative, qualitative, and mixed-methods) to assess both implementation processes and outcomes. Finally, it endeavors to integrate advances in developmental science to inform more precise targeting of program components.

The CIRCLE approach is not intended to replace existing evaluation methods but rather to complement them with a more flexible, nuanced perspective. It recognizes the value of diverse forms of evidence while maintaining a commitment to rigorous testing of causal hypotheses. By embracing methodological pluralism, prioritizing understanding of impact variation, and emphasizing precision in theory and measurement, this approach has the potential to advance the field of early childhood program evaluation toward more accurate, more equitable, and ultimately more impactful interventions.

The CIRCLE Framework in Practice: A Six-Step Process for Continuous Learning and Improvement

While the theoretical foundations of the CIRCLE framework provide a conceptual roadmap for transforming early childhood program evaluation, the practical implementation of this approach requires a systematic process that guides programs through cycles of learning and refinement. The CIRCLE approach consists of six interconnected steps that operationalize the framework's core principles of continuous quality improvement, understanding impact variation, and precision in theory and measurement.

Step 1: Developing a Theory of Impact



The CIRCLE approach begins with the development of a detailed theory of impact (TOI), which serves as the foundation for all subsequent evaluation activities (see template in Figure 3). Unlike traditional logic models or theories of change that often focus primarily on outputs and outcomes,

the TOI provides a more nuanced framework that distinguishes among four key components—values, strategies, targets, and outcomes—while also identifying potential moderators of impact.

The TOI development process typically involves three workshops with program stakeholders. The first workshop introduces partners to the CIRCLE framework's continuous improvement mindset, emphasizing learning about it over proving program effectiveness. This session begins

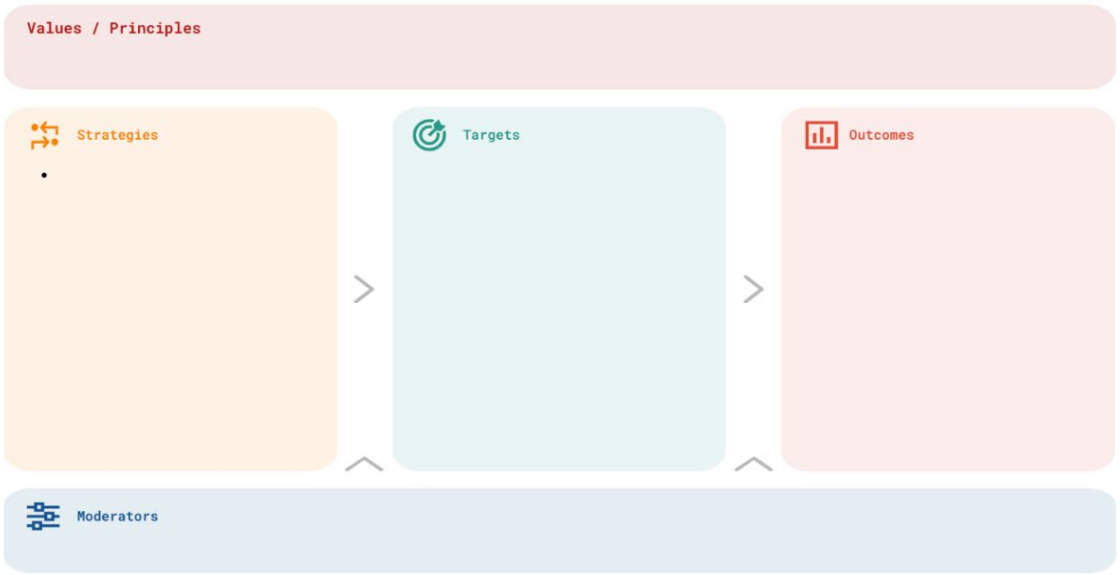


Figure 3. Theory of Impact template

by exploring the program's underlying values—the fundamental "why" that drives the work. Understanding these values is critical because they ground all subsequent discussions and ensure that evaluation activities align with the program's core mission and commitments.

Following the values exploration, the first workshop focuses on identifying program strategies—the specific activities and interventions that comprise the program. Rather than simply listing activities, this process encourages programs to think carefully about what they actually do, considering both the sequence of activities and how they might be grouped by delivery mechanism or target population. This detailed examination of strategies often reveals assumptions about program implementation that may need to be tested or refined.



The second workshop shifts focus to targets—the immediate, proximal impacts that programs expect to see as a direct result of their strategies. This represents a critical distinction from traditional evaluation approaches that often jump directly from activities to long-term outcomes. By identifying targets, programs develop clearer hypotheses about the causal mechanisms through which their strategies are expected to work. For example, a parent education program might identify "increased parental confidence" as a target that mediates the relationship between home visiting strategies and longer-term child development outcomes.

The third workshop addresses outcomes—the more distal goals that programs hope to achieve. These might include changes in child development, family functioning, or even systems-level impacts. Importantly, the TOI process recognizes that outcomes may extend beyond individual participants to include broader community or policy changes. This workshop also identifies moderators—factors that may influence the extent to which different participants benefit from the program. These might include participant characteristics (such as prior experience with similar interventions), program features (such as dosage or language of delivery), or contextual factors (such as community resources).

The completed TOI serves multiple purposes beyond evaluation planning. It provides a shared language and understanding among stakeholders, reveals areas where program theory may need refinement, and identifies assumptions that require testing. Importantly, the TOI is treated as a living document that can be updated as programs learn and evolve.

Step 2: Selecting Appropriate Measures

Once the TOI is complete, the next step involves selecting measures that can capture each component of the theory. This process operationalizes the CIRCLE framework's emphasis on precision in measurement by ensuring that assessment strategies are directly aligned with program targets and expected outcomes.

The measure selection process begins with determining which aspects of the TOI are most critical to measure given the program's stage of development and primary learning questions. For newer programs, it may be essential to focus initially on measuring strategies to ensure fidelity and confirm that the program is operating as intended. More established programs might prioritize measuring targets to understand whether the hypothesized immediate impacts are occurring or outcomes to assess longer-term effectiveness.

Measure selection involves balancing several considerations. When validated measures exist for key constructs, they are generally preferred because they allow for comparison across studies and contexts. However, practical constraints such as survey length, participant burden, or cultural appropriateness may necessitate the development of new measures or adaptation of existing ones. The CIRCLE approach recognizes that perfect measurement is often the enemy of useful measurement and encourages programs to select measures that are "good enough" to support learning while remaining feasible to implement.

This step also involves initial consideration of the broader measurement strategy, including the balance between quantitative and qualitative approaches, the frequency of data collection, and the methods for collecting information from different stakeholder groups, to be further expanded upon in step 3. The goal is to develop a measurement plan that provides sufficient information to support continuous learning while remaining sustainable for the program and minimally burdensome for participants.

Step 3: Determining Research Design



With measures selected, programs can develop a comprehensive research design that specifies how data will be collected to answer key learning questions. This step embodies the CIRCLE framework's commitment to methodological pluralism by calibrating the rigor of evaluation to the program's developmental stage and specific learning needs.

The research design process deeply considers several key factors: timing of data collection, groups to be sampled, available data sources, and the balance between quantitative and qualitative methods. For programs in early stages of development, the design might emphasize implementation monitoring and rapid-cycle feedback to support program refinement. More established programs might employ quasi-experimental designs or, when appropriate and feasible, RCTs to test specific causal hypotheses.

Importantly, the CIRCLE approach does not default to RCTs as the gold standard for all evaluation questions. Instead, it recognizes that RCTs, while valuable for establishing causal

relationships, have significant limitations in terms of cost, time, equity considerations, and contextual relevance. For many programs, particularly those serving historically marginalized populations, withholding potentially beneficial interventions from control group participants may raise serious ethical concerns. The framework instead supports a continuum of evidence, where different research designs are matched to specific evaluation questions and contexts.

Alternative approaches such as pre-post studies, time series designs, or comparison group studies may be more appropriate for many evaluation questions, particularly those focused on understanding implementation processes, identifying differential impacts across subgroups, or exploring mechanisms of change. The key is selecting a design that provides sufficient rigor to support confident conclusions while remaining feasible and ethical within the program's context.

Step 4: Collecting Data



Data collection strategies are matched to the selected measures and research design, with careful attention to ensuring that data collection processes are culturally responsive and minimize burdens on participants and program staff. The CIRCLE approach emphasizes the importance of community input in designing data collection procedures, recognizing that community members are best positioned to understand what approaches will work in their specific contexts.

Data collection may involve quantitative surveys, qualitative interviews or focus groups, implementation monitoring tools, or mixed-methods approaches that combine multiple data sources. The framework's emphasis on rapid-cycle learning often involves more frequent data collection than traditional evaluation approaches, allowing programs to make timely adjustments based on emerging findings.

Step 5: Analyzing Data and Sharing Learning



The analysis phase employs appropriate methodological approaches based on the type of data collected and the specific learning questions being addressed. Consistent with the CIRCLE framework's emphasis on understanding impact variation, analyses routinely examine differential impacts across subgroups rather than focusing solely on average treatment effects.

A distinctive feature of the CIRCLE approach is its commitment to rapid sharing of findings to support



continuous learning. Rather than waiting for final reports, the process emphasizes creating data slides and other accessible formats that can be used to facilitate discussions with partners, participants, and other stakeholders. These conversations serve multiple purposes: they help with interpretation of findings, inform adaptations to measures or implementation, and prioritize future analyses.

The framework's commitment to equity extends to knowledge sharing, ensuring that findings are communicated back to participants and community members in accessible formats. This might involve data sections in community newsletters, text message summaries, or community meetings to discuss results and implications.

Step 6: Reflecting and Incorporating Learning



The final step in each cycle of the CIRCLE approach involves pausing to reflect on what has been learned and how this learning should inform future program development and evaluation activities. This reflection process examines not only the substantive findings about program impacts but also lessons about the evaluation process itself—what worked well, what could be improved, and what questions emerged that warrant further investigation.

Based on this reflection, the TOI may be updated to reflect new understanding about program mechanisms, revised theories about how change occurs, or shifts in program priorities. These updates then inform the next cycle of evaluation activities, embodying the framework's commitment to continuous learning and improvement.

This systematic yet flexible process operationalizes the CIRCLE framework's vision of evaluation as a tool for ongoing learning rather than a one-time judgment about program effectiveness. By implementing this process, early childhood programs can build their capacity for continuous improvement while contributing to the broader field's understanding of effective intervention approaches. The process recognizes that achieving meaningful impact requires ongoing adaptation and refinement, supported by systematic inquiry that is both rigorous and responsive to community needs and contexts.



Implications of the CIRCLE Framework for Policy, Practice, and Research

Implications for Policy

For policy makers, the CIRCLE approach suggests a fundamental reconsideration of how evidence is used in funding and program decisions. Rather than relying solely on evidence from RCTs to identify "what works," policy makers should embrace a broader understanding of evidence that provides insights into how programs work, for whom, and under what conditions.

This shift requires expanding definitions of evidence to encompass a continuum of research designs, with the rigor of evidence calibrated to the stage of program development and the specific questions being addressed. Policy makers should prioritize funding for programs that demonstrate a genuine commitment to understanding and addressing impact variation, particularly for underserved populations who have historically been marginalized in intervention research. Supporting the development of shared measurement systems is another critical policy priority, enabling comparison across programs and contexts while maintaining sensitivity to program-specific targets and local circumstances.

Investment in research infrastructure that facilitates rapid-cycle evaluation and continuous quality improvement should take precedence over exclusively supporting definitive impact studies.

This approach recognizes that sustainable change requires ongoing learning and adaptation rather than single, high-stakes evaluations that may not capture the complexity of real-world implementation.

Implications for Programs and Practice

For programs and practitioners, the CIRCLE approach presents both significant challenges and unprecedented opportunities. The framework's emphasis on precision in theory and measurement demands greater clarity about program mechanisms while its focus on impact variation requires more sophisticated attention to how programs might be adapted for different populations and contexts.

Programs must develop more detailed theories of change that explicitly specify the causal chains linking program activities to expected outcomes, with particular attention to the proximal targets that programs directly influence. This level of specificity moves beyond traditional logic models to create actionable frameworks for understanding how change occurs. Incorporating systematic assessment of implementation quality and fidelity should become a

routine part of program operation rather than an afterthought, providing ongoing feedback about whether programs are being delivered as intended.

Perhaps most important, programs should use data on patterns of differential response to identify subgroups that may benefit less from current approaches and develop targeted adaptations accordingly. This requires moving beyond one-size-fits-all models toward more nuanced, responsive programming.

The ultimate goal for programs is engaging in continuous learning through rapid cycles of innovation, testing, and refinement rather than waiting for definitive evidence of effectiveness before making improvements.

Implications for Research

For researchers, the CIRCLE approach calls for greater methodological pluralism and a fundamental shift in focus from average effects to patterns of differential impact. This transformation also suggests the need for closer integration between developmental science and intervention research, ensuring that evaluation approaches are informed by the latest understanding of how children develop and learn.

Researchers should employ a broader range of evaluation designs, tailoring specific approaches to the stage of program development and the particular questions being addressed. This methodological flexibility allows for more appropriate matching of research methods to research questions. Examining and reporting patterns of differential impact should become routine practice, encompassing both prespecified moderators based on theory and exploratory analyses designed to identify unexpected patterns of response.

The field needs continued development and refinement of methods for assessing individual-level change and relating these patterns to group-level effects, moving beyond statistical significance to understand practical significance for individual participants.

Cross-disciplinary collaboration becomes essential to ensure that intervention approaches are informed by advances in developmental science, neuroscience, and related fields. Researchers should also explore innovative approaches to rapid-cycle evaluation that can provide timely feedback for program refinement while maintaining appropriate scientific rigor for the questions being asked.

Reimagining Incentives for Learning and Improvement

For Academic and Research Institutions

Academic institutions must fundamentally broaden their definition of scholarly impact to value contributions to practice and policy alongside traditional academic metrics. This transformation involves rewarding publication of null findings, failed interventions, and studies that illuminate differential impacts or implementation challenges rather than privileging only positive results. Partnerships with practitioners and community organizations should be valued in tenure and promotion decisions, recognizing that meaningful research often emerges from collaborative relationships with those implementing and receiving interventions.

Creating dedicated funding streams for rapid-cycle evaluation, replication studies, and research on implementation processes represents a critical institutional priority. Academic institutions should also develop new outlets for sharing "lessons learned" that may not meet traditional publication thresholds but provide valuable insights for practice. As Fisher and colleagues (2020) suggest, the academic community must embrace a culture of "short-cycle sharing" that promotes quick turnaround of information, encourages rapid discovery, and values learning from failure. This cultural shift requires new publication formats, conference structures, and collaboration mechanisms that facilitate timely exchange of emerging findings.

For Programs and Practitioners

Programs need restructured incentives to engage honestly in evaluation and use findings for improvement rather than merely compliance. This includes providing dedicated funding for internal evaluation capacity that focuses on learning and adaptation rather than exclusively on accountability measures. Programs should receive protection from funding penalties when rigorous evaluation reveals implementation challenges or limited impacts, especially when these findings are accompanied by concrete plans for improvement and continued learning.

Recognition and rewards should emphasize transparency about challenges and creative problem solving rather than focusing solely on positive outcomes. Establishing peer learning communities that normalize sharing of struggles and collective problem solving can create supportive environments for honest reflection and improvement. The experience of continuous quality improvement in health care provides instructive parallels—when providers are protected from punitive consequences for identifying problems and instead rewarded for transparent reporting and systematic improvement efforts, quality improves more rapidly than in punitive environments (Berwick et al., 2003).

For Federal and Philanthropic Funders

Funders possess perhaps the greatest power to reshape incentives through their grant-making practices and should embrace this responsibility. Grant structures should include distinct

phases for development, refinement, and testing, with appropriate evaluation approaches tailored to each phase rather than applying uniform standards across different stages of program maturation. A habit of requiring and funding robust implementation monitoring alongside outcome evaluation supports understanding of how programs achieve their effects rather than simply whether they work.

Creating "learning portfolios" rather than funding isolated programs allows lessons from multiple interventions to be synthesized for building collective knowledge. Investment in evaluation infrastructure that supports rapid-cycle methods and shared measurement across multiple programs can accelerate learning across the field. Providing longer funding horizons enables iterative refinement based on evaluation findings rather than forcing programs into artificial timelines that may not align with the realities of sustainable change.

CIRCLE Case Example: A Reimagined Approach

To illustrate how transformed incentives might work in practice, consider the Talk Together program, a family-centered early language intervention designed to enhance language development among children from birth to three years in low-income communities. The program combines weekly home visits with monthly group sessions, focusing on training caregivers to use responsive interaction techniques, shared book reading strategies, and everyday conversation enhancement methods. Home visitors, who are bilingual community members with early childhood training, work with families to identify daily routines—such as mealtime, bath time, and grocery shopping—as opportunities for rich language interaction. The program also provides families with carefully curated books and simple learning materials that align with children's developmental stages.

Within the current early childhood program evaluation ecosystem, Talk Together might undergo a single, high-stakes RCT after several years of implementation, with results determining whether funding continues—a binary approach that provides limited learning opportunities and fails to capture the program's complexity across diverse community contexts.

Under a reimagined approach with learning-oriented incentives, Talk Together would begin with a detailed theory of impact specifying how home visiting strategies and group sessions are expected to increase caregiver confidence and responsive interaction skills (targets), which in turn should improve children's vocabulary development and communication abilities (outcomes). The theory would also identify potential moderators such as caregiver education level, primary language spoken at home, and frequency of program participation.

Early implementation would be accompanied by rapid-cycle testing of key program components across different sites. For example, researchers might examine whether monthly group sessions are more effective when focused on specific skills (like shared reading) versus when they address broader topics (like supporting child development). They might test different approaches to training home visitors or explore how program materials can be better adapted for families who speak languages other than English or Spanish. Findings from these



rapid cycles would continuously inform program refinement—perhaps leading to more frequent group sessions in some communities, modified training protocols for home visitors, or culturally adapted materials for specific populations.

As implementation stabilized across multiple sites, more rigorous evaluation would examine not only average effects on child language outcomes but also variation across different types of families and communities. Researchers might discover that the program is particularly effective for first-time parents but shows smaller impacts for families with multiple children, or that certain components work better in urban versus rural settings. These findings would lead to targeted adaptations rather than abandonment of the entire approach.

Throughout this process, researchers would receive recognition for identifying both effective components and implementation challenges, with findings published regardless of whether overall impacts were positive, contributing to honest knowledge building about early language intervention. When rapid-cycle testing revealed that group sessions were poorly attended in certain communities, researchers would be rewarded for documenting this challenge and exploring alternative approaches rather than hiding negative findings.

Funders would support ongoing refinement based on evaluation findings rather than making binary decisions about continuation. When data showed that Talk Together was less effective for families experiencing housing instability, additional funding might support development of modified protocols that address the unique needs of highly mobile families rather than terminating the program entirely.

Talk Together would participate in a learning community with similar language-focused interventions, sharing implementation strategies, comparing outcomes across different approaches, and collectively solving challenges like staff retention or family engagement. This collaboration would accelerate learning beyond what any single program could achieve in isolation.

Policy decisions about early language programming would be based on sophisticated understanding of what specific components work, for which types of families, and under what implementation conditions, rather than simplified ratings of "evidence-based" status that treat Talk Together as either effective or ineffective. Policy makers might learn, for example, that home visits combined with group sessions are most effective for engaging families but that the specific content and frequency should vary based on community characteristics and family needs.

This comprehensive approach would generate substantially more useful knowledge for the field of early language intervention, support continuous improvement of Talk Together across diverse implementation sites, and ultimately lead to more effective, culturally responsive interventions that better serve children and families in their specific contexts.



Conclusion

The field of early childhood intervention stands at a crossroads. Despite decades of research demonstrating the value of investing in the early years, we have yet to achieve the kind of population-level impacts that early pioneers in the field envisioned. Moving forward requires a critical examination of how we evaluate early childhood programs and translate findings into practice.

The challenges identified in this paper—limitations of the RCT paradigm, overemphasis on average treatment effects, lack of precision in theory and measurement, and misaligned incentive structures—are not insurmountable. By embracing a more nuanced approach to evaluation that recognizes the value of diverse forms of evidence, prioritizes understanding of impact variation, emphasizes greater precision in theory and measurement, and transforms incentives to support learning, we can advance the field toward more effective interventions and, in turn, broader impacts.

The CIRCLE framework proposed here offers a starting point for this evolution. By calibrating evaluation approaches to the stage of program development, systematically examining patterns of differential impact, and drawing on advances in developmental science to inform more precise targeting of program components, we can build on the substantial achievements of the past half century while addressing persistent challenges.

The ultimate goal remains the same as that articulated by early pioneers in the field: to develop effective interventions that improve outcomes for all children, particularly those facing significant adversity. Achieving this goal requires not only continued investment in early childhood programs but also a commitment to innovation in how we design, implement, and evaluate these programs. By embracing this challenge, we can work toward a future in which all children have the opportunity to reach their full potential.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Belfield, C. R., Nores, M., Barnett, W. S., & Schweinhart, L. (2006). The High/Scope Perry Preschool Program: Cost–benefit analysis using data from the age-40 followup. *Journal of Human Resources, 41*(1), 162–190.

Berwick, D. M., James, B., & Coye, M. J. (2003). Connections between quality measurement and improvement. *Medical Care, 41*(1), I-30–I-38.

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., ... & Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology, 48*(4), 1033–1043.

Cook, T. D. (2018). Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: A commentary on Deaton and Cartwright. *Social Science & Medicine, 210*, 37–40.

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal, 44*(2), 298–339.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109–132.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350.

Estrada, E., Ferrer, E., & Pardo, A. (2019). Statistics for evaluating pre-post change: Relation between change in the distribution center and change in the individual scores. *Frontiers in Psychology, 9*, 2696.

Fisher, P. A., Barker, T. V., & Blaisdell, K. N. (2020). A glass half full and half empty: The state of the science in early childhood prevention and intervention research. *Annual Review of Developmental Psychology, 2*, 269–294.

Grossman, J., & Mackenzie, F. J. (2005). The randomized controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine, 48*(4), 516–534.

Hahn, R. A., & Barnett, W. S. (2023). Early childhood education: Health, equity, and economics. *Annual Review of Public Health*, *44*(1), 75–92.

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology, 27*(5), 711–731.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the High/Scope Perry Preschool Program. *Journal of Public Economics, 94*(1), 114–128.

Horn, S. R., Fisher, P. A., Pfeifer, J. H., Allen, N. B., & Berkman, E. T. (2020). Levers and barriers to success in the use of translational neuroscience for the prevention and treatment of mental health and promotion of well-being across the lifespan. *Journal of Abnormal Psychology*,

Jacobson, N. S., & Truax, P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives, 25*(3), 17–38.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., ... & Downer, J. (2012). Third grade follow-up to the Head Start Impact Study: Final report. OPRE Report 2012-45, Administration for Children & Families.

Raudenbush, S. W. (2018). On randomized experimentation in education: A commentary on Deaton and Cartwright, in honor of Frederick Mosteller. *Social Science & Medicine, 210*, 63–66.

Schindler, H. S., McCoy, D. C., Fisher, P. A., & Shonkoff, J. P. (2019). A historical look at theories of change in early childhood education research. *Early Childhood Research Quarterly, 48*, 146–154.

Schorr, L. B., & Farrow, F. (2011). Expanding the evidence universe: Doing better by knowing more. Center for the Study of Social Policy.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry Preschool study through age 40. HighScope Press.

Shonkoff, J. P., & Fisher, P. A. (2013). Rethinking evidence-based practice and two-generation programs to create the future of early childhood policy. *Development and Psychopathology, 25*(4pt2), 1635–1653.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., ... & Gershon, R. (2013). The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society, 19*(6), 671–681.

What Works Clearinghouse. (2020). What Works Clearinghouse Standards Handbook, Version 4.1. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Whitaker, A. A., Burchinal, M., Jenkins, J. M., Bailey, D. H., Watts, T. W., Duncan, G. J., ... & Peisner-Feinberg, E. (2023). Why are preschool programs becoming less effective?. *Journal of Policy Analysis and Management*.