Program Evaluation and Implementation Research in Education: An Evidence-Informed Framework for Measuring Impact and Scaling Effective Practices

Javier Solano Pineda
Universidad Tecnológica de Pereira, Colombia
Corresponding author: jsolano@utp.edu.co

Abstract

Educational innovations across diverse national and institutional contexts routinely diffuse and scale before they have been adequately evaluated, producing costly, system-wide reforms whose effects on student learning remain ambiguous and whose equity consequences frequently go unexamined until disparities are already entrenched. Simultaneously, traditional evaluation approaches that prioritize summative impact verdicts over explanatory insight often fail to illuminate why an intervention produces strong outcomes in one implementation context while underperforming in another, leaving practitioners and policymakers unable to distinguish design failures from implementation failures or to identify the contextual conditions that determine whether an intervention's theoretical mechanisms actually operate. This evidence-informed conceptual paper synthesizes program evaluation traditions, including logic models, theory of change methodology, and utilization-focused and developmental evaluation approaches, with implementation research constructs of fidelity, adaptation, reach, and feasibility, and with contemporary principles of impact measurement and equity-sensitive indicator design, to propose a practical framework for learning-oriented evaluation of educational innovations. Drawing on widely used frameworks including RE-AIM, the Consolidated Framework for Implementation Research, and principles of construct-valid outcome measurement, the paper articulates four interdependent domains: (a) a clearly specified theory of change that distinguishes core intervention components from adaptable elements and explicitly names mediating mechanisms; (b) implementation measurement with equity-sensitive indicators and interpretive guardrails that protect data from punitive misuse; (c) outcome measurement grounded in construct validity and triangulated across multiple evidence sources; and (d) scaling decisions structured as staged, evidence-guided learning processes rather than as threshold-based deployment verdicts. Three conceptual tables operationalize the framework through a theory-of-change template for educational interventions, an implementation and outcome indicator menu with interpretation guardrails, and a scaling readiness checklist for institutional leaders. The paper concludes with recommendations for evaluators, researchers, funders, and education system leaders seeking to measure impact responsibly while accelerating the improvement of educational practice at scale.

Keywords: *Program Evaluation; Implementation Research; Impact Measurement; Scaling; Education Reform; Theory Of Change; Equity-Sensitive Evaluation.*

A. INTRODUCTION

The relationship between educational innovation and evidence has long been characterized by a structural asymmetry that imposes significant costs on education systems and the students they serve. Innovations spread through policy mandates, funding incentives, peer diffusion, and professional enthusiasm, carried by plausible theories and compelling early-adopter accounts, while the evaluation infrastructure needed to generate credible, timely, and contextually informative evidence about their effects lags far behind the pace of adoption. By the time rigorous evidence about an intervention's impact accumulates in the research literature, the intervention has often already been mandated, scaled, embedded in institutional practice, and defended by the professional identities and political commitments of those who championed its adoption. The subsequent discovery that effects are smaller than anticipated, unevenly distributed across student populations, or dependent on implementation conditions that were not present in many adopting sites then produces the familiar and costly cycle of reform, reaction, and abandonment that has characterized educational policy in many systems across multiple generations of innovation.

This pattern is not merely unfortunate; it is structurally predictable given the incentive architecture within which educational innovation typically operates. Funders and policymakers who are accountable for demonstrating action on educational challenges have strong incentives to support visible, rapidly diffusing innovations and weak incentives to invest in the slow, methodologically demanding

process of understanding how and why those innovations produce or fail to produce their intended effects. Program developers who have invested years in creating and refining an intervention have strong incentives to demonstrate its effectiveness and weak incentives to generate the explanatory evidence that might reveal the conditions under which it fails. Practitioners who have committed professional effort to implementing an innovation have strong incentives to report positive experiences and weak incentives to document implementation challenges that might be interpreted as personal failures. The cumulative effect of these incentive structures is an evidence ecosystem characterized by optimistic early studies, slow-accumulating disconfirmation, and recurring debates about whether disappointing results reflect the inherent limitations of an idea or the preventable failures of its execution.

Program evaluation and implementation research represent complementary responses to this structural problem, offering conceptual tools and methodological approaches that can reduce the gap between innovation adoption and evidence quality. Program evaluation traditions, spanning the logic model approaches that make program theories explicit, the theory of change methodologies that trace causal pathways from activities to outcomes, and the utilization-focused and developmental evaluation orientations that embed evidence generation in ongoing improvement processes, provide frameworks for asking and answering the right questions about whether and how interventions produce their intended effects. Implementation research, drawing on health sciences traditions that have been productively applied to educational contexts, provides constructs and measures for understanding the delivery conditions, organizational factors, and adaptation processes that mediate between intervention design and outcome achievement, making it possible to distinguish what an intervention does under optimal implementation conditions from what it delivers in the diverse and often sub-optimal conditions of real educational systems.

The integration of these two traditions within a learning-oriented evaluation framework is the central contribution of this paper. Learning-oriented evaluation, as the paper develops the concept, is distinguished from compliance-oriented evaluation not primarily by its methods but by its purposes and its epistemological orientation: it treats evaluation as a governance and improvement function whose central purpose is to generate the contextually sensitive, actionable, and equity-attentive evidence that educational decision-makers at all levels need to make better choices about which innovations to adopt, how to implement them, when to adapt them, and when to discontinue them. This orientation does not require sacrificing rigor for relevance; on the contrary, it demands a more sophisticated conception of rigor than simple effect size estimation provides, one that attends to construct validity, equity implications, implementation fidelity and adaptation, and the contextual conditions that moderate intervention effects.

The paper proceeds as follows. The subsequent section develops the theoretical and empirical foundations of the framework through a structured review of program evaluation traditions, implementation research concepts, and equity-sensitive impact measurement principles. The third section presents the four-domain framework with three operationalizing tables designed for practitioner use. The fourth section addresses common evaluation pitfalls, design implications for leaders and funders, and the organizational conditions that support learning-oriented evaluation cultures. The concluding paragraph synthesizes the framework's contributions and its central practical message for educational innovators and evaluators.

B. LITERATURE REVIEW

Program Evaluation Traditions: From Logic Models to Developmental Evaluation

The intellectual history of program evaluation in education reflects a progressive elaboration of the fundamental question: how do we know whether an educational intervention is working, for whom, and why? Early evaluation traditions, associated with the experimental and quasi-experimental designs that dominated educational research in the 1960s and 1970s, prioritized causal attribution through rigorous research designs that isolated intervention effects from confounding variables. This tradition produced important methodological advances in identifying average treatment effects but was limited in its capacity to explain the mechanisms through which effects were produced, the conditions under which they varied, or the equity implications of interventions that might produce positive average effects while widening disparities across student subgroups.

The development of logic models and theory of change methodology represented a significant advance in evaluation practice by making program theories explicit and testable. Logic models, which map the causal chain from program inputs and activities through intermediate outcomes to long-term goals, provide a structured basis for evaluation planning by identifying the specific mechanisms and outcomes that evaluation should measure and the assumptions about context and capacity whose validity the evaluation should examine (Rogers, 2008). Theory of change methodology extends this approach by

requiring program developers and evaluators to articulate the causal logic through which program activities are expected to produce outcomes, including the intermediate steps, enabling conditions, and boundary conditions that the causal pathway depends on. When evaluation is designed to test a theory of change rather than simply to measure outcomes, it generates explanatory as well as descriptive knowledge: not only whether the program worked but why it worked or failed to work, which conditions were necessary for its mechanisms to operate, and which elements of the program's theory require revision in light of evidence.

Utilization-focused evaluation, developed by Patton (2008) as a response to the well-documented tendency of evaluation findings to be ignored by the decision-makers they were intended to inform, reorients evaluation design around the practical decisions that stakeholders actually face rather than around the methodological standards of the research community. A utilization-focused evaluation begins by identifying the specific decisions that evaluation evidence will need to inform, the decision-makers who will use that evidence, and the forms of evidence those decision-makers find credible and actionable. This stakeholder orientation does not compromise methodological quality but ensures that methodological choices serve practical purposes: a randomized controlled trial that produces generalizable causal estimates three years after a program has already scaled is methodologically rigorous but pragmatically useless; a mixed-methods evaluation that generates timely, contextually rich evidence about implementation quality and outcome patterns may produce less precise causal estimates while substantially improving the quality of ongoing implementation decisions.

Developmental evaluation, conceptualized by Patton as an adaptation of utilization-focused evaluation for complex, dynamic innovations, goes further by positioning the evaluator as an embedded partner in an innovation process rather than an external assessor of a stable program. In developmental evaluation, evidence generation and program development are simultaneous and mutually informing processes: evaluation data continuously informs adaptation decisions, and each adaptation cycle generates new evidence about what works under what conditions. This orientation is particularly appropriate for educational innovations that are inherently context-dependent and that require ongoing adaptation to be effective across diverse implementation sites, but it requires a level of evaluator-innovator collaboration and organizational trust that many evaluation contexts do not currently support.

Implementation Research: Fidelity, Adaptation, and the Context-Dependence of Impact

Implementation research emerged from a recognition, crystallized in Fixsen and colleagues' (2005) influential synthesis, that the gap between an intervention's demonstrated efficacy under optimal conditions and its achieved effectiveness in real-world delivery contexts is not primarily a product of insufficient knowledge about what works but of insufficient understanding of the implementation processes and organizational conditions through which evidence-based practices are or are not faithfully and skillfully delivered. The practical implication is consequential: investing in the development of more effective interventions without simultaneously investing in the implementation science needed to deliver existing effective interventions with quality and consistency may produce diminishing returns on the research investment.

The Consolidated Framework for Implementation Research provides the most comprehensive conceptual map of the factors that shape implementation quality and outcomes in complex organizational settings (Damschroder et al., 2009). Organized across five domains, the CFIR identifies implementation determinants in the characteristics of the intervention itself, the inner organizational setting in which implementation occurs, the outer setting of policy and professional context, the characteristics of the individuals involved in implementation, and the process by which implementation is planned and managed. In educational contexts, these domains translate into a recognizable set of implementation determinants: the degree to which an intervention fits existing curriculum and assessment structures; the quality and consistency of the professional learning support provided to implementing teachers; the leadership routines that protect implementation time and signal institutional priority; the policy environment that creates incentives or constraints for adoption; and the teacher-level factors of knowledge, motivation, and professional identity that shape how interventions are interpreted and enacted in classroom practice.

The constructs of fidelity and adaptation occupy a conceptually important and practically contested position in implementation research. Fidelity, defined as the degree to which an intervention is implemented as designed, has historically been treated as a quality indicator: higher fidelity is assumed to produce better outcomes because it ensures that the theory of change operates as intended. Research on this assumption has produced a more nuanced understanding: fidelity to the core components of an intervention, those elements whose integrity is necessary for the theory of change to function, is associated with better outcomes, but rigid adherence to peripheral or structural features of an

intervention that were not central to its theory of change frequently undermines implementation quality by preventing the adaptations that local context requires (Durlak and DuPre, 2008). The distinction between core and adaptable components is therefore not merely a theoretical nicety but a practically consequential design decision that evaluation must track: evaluation should monitor fidelity to theoretically essential components while documenting and analyzing adaptations to understand which modifications strengthen implementation in context and which undermine the intervention's mechanism.

The RE-AIM framework, originally developed by Glasgow, Vogt, and Boles (1999) for public health intervention evaluation, has been productively applied to educational contexts as a multidimensional evaluation rubric that extends beyond efficacy to address the full range of factors that determine whether an intervention can produce population-level impact. RE-AIM evaluates interventions across five dimensions: Reach, the proportion of the intended population that participates; Effectiveness, the magnitude and breadth of outcomes across participants; Adoption, the proportion of eligible settings that implement the intervention; Implementation, the consistency and quality of delivery; and Maintenance, the degree to which intervention effects and implementation practices are sustained over time. This multidimensional framework makes visible the distinction between an intervention that produces strong effects in the settings that adopt it enthusiastically and one that produces population-level impact by achieving effects of adequate magnitude across a broadly representative reach of the intended population, a distinction that is systematically obscured by effect-size-focused evaluations that do not attend to reach and adoption.

Impact Measurement, Construct Validity, and the Equity Imperative

The measurement of educational outcomes in program evaluation contexts raises fundamental questions of construct validity that are frequently underattended in evaluation practice. Construct validity, in Messick's (1995) comprehensive formulation, refers to the degree to which the inferences drawn from assessment or measurement results are appropriate, meaningful, and defensible given the constructs those measures are intended to represent. In educational evaluation, construct validity failures occur when outcome measures assess constructs that are correlated with but importantly different from the outcomes an intervention is designed to improve: a digital learning intervention designed to improve critical thinking that is evaluated using a multiple-choice recall test may show no effect not because it failed to improve critical thinking but because the outcome measure did not assess the construct the intervention targeted.

The proliferation of digital learning environments has created a particular measurement risk that deserves explicit attention: the availability of abundant behavioral trace data, including login frequency, time-on-task, content completion rates, and discussion post counts, creates the temptation to substitute easily collected proxy indicators for the harder-to-measure learning constructs that interventions are designed to improve. These proxy indicators are not without evaluative value; they can provide useful signals of engagement patterns and implementation reach, and they can serve as early warning indicators of disengagement risk. Used as primary outcome measures in place of valid learning assessments, they produce misleading evaluative conclusions: an intervention that increases login frequency and content completion rates while producing no improvement in students' capacity to reason, analyze, or apply knowledge will appear effective by proxy indicator standards while failing by the standards that matter educationally.

Equity-sensitive evaluation requires disaggregating both implementation and outcome data across student subgroups defined by characteristics such as socioeconomic status, race and ethnicity, gender, disability status, language background, and geographic location, attending to whether interventions produce benefits that are equitably distributed across their intended population or whether positive average effects conceal divergent patterns in which some subgroups benefit substantially while others are unaffected or actively harmed. The equity evaluation question is not simply whether an intervention works on average but whether it works for the specific student populations whose improvement is most urgently needed and whose experiences are most frequently obscured by aggregate reporting. Disaggregated evaluation analysis requires careful attention to the ethical conditions of data governance: collecting, storing, and reporting data in ways that could expose individual students or identifiable subgroups to harm requires consent-based data collection, secure storage, restricted access, and interpretive guardrails that prevent disaggregated data from being used to deficit-frame the communities whose experiences it documents.

C. METHOD

This paper employs an evidence-informed conceptual framework development methodology, constructing a theoretical model through systematic synthesis of established research traditions in

program evaluation, implementation science, and impact measurement rather than through original empirical data collection. This approach is consistent with scholarly practice in educational administration, evaluation methodology, and implementation research, where conceptual framework papers serve the important function of integrating distributed literatures into actionable models that can guide both research and practice.

The literature synthesis proceeded in three stages. In the first stage, foundational theoretical and methodological works in the primary scholarly traditions informing the framework were identified through systematic searches of major educational and social science research databases, including ERIC, PsycINFO, the Web of Science, and Google Scholar. Search terms spanned program evaluation, theory of change, logic model, implementation fidelity, implementation research, implementation science, RE-AIM, CFIR, scaling educational innovations, impact measurement, construct validity, and equity-sensitive evaluation. Priority was accorded to peer-reviewed journal articles, scholarly monographs, and research synthesis reports from established research institutions, with particular attention to works that have achieved broad adoption in evaluation practice communities. In the second stage, empirical literature from the preceding fifteen years was reviewed to identify findings that extend, qualify, or update the foundational frameworks in light of contemporary educational reform and evaluation contexts, with emphasis on studies conducted in large-scale educational evaluation settings that illuminate the practical implementation and organizational dynamics of learning-oriented evaluation. In the third stage, cross-domain synthesis was conducted to identify the theoretical relationships among the program evaluation, implementation research, and impact measurement traditions that the framework integrates, mapping their complementary contributions and productive tensions onto the four-domain architecture of the proposed model.

Three conceptual tables were constructed as operationalizing instruments for the framework, each designed to translate the framework's theoretical principles into specific design choices and decision criteria for practitioner use. These tables were developed through iterative refinement against the framework's internal logic and the practical constraints of evaluation design in resource-variable educational contexts. As a conceptual framework paper, the propositions advanced here are theoretical rather than empirically validated through original research, and the framework's credibility rests on the coherence of its integration and the quality of the evidence base from which it is constructed. Empirical examination of the framework's domain relationships across diverse educational evaluation contexts is identified as a priority for future research.

D. RESULT AND DISCUSSION

Building a Clear Theory of Change and Specifying Core Components

The first domain establishes the evaluative foundation on which all subsequent measurement and interpretation depends. A theory of change, as the framework employs the concept, is not merely a visual map of program logic but a structured causal argument that specifies the mechanisms through which program activities are expected to produce outcomes, the intermediate steps and mediating variables through which those mechanisms operate, the enabling conditions that must be present for the mechanisms to function, and the boundary conditions that limit the contexts in which the theory is expected to hold. This level of specificity serves evaluation in three ways: it identifies the specific mediators and outcomes that evaluation should measure to test whether the theory is operating as predicted; it makes the contextual conditions whose presence the theory assumes into testable propositions that evaluation can examine rather than assumptions that evaluation ignores; and it creates a shared understanding between program developers, evaluators, and practitioners about what constitutes the essential core of an intervention as distinguished from the contextually variable elements that implementation must adapt to local conditions.

The distinction between core and adaptable components deserves particular emphasis as a practical evaluation design decision. Core components are those whose integrity is necessary for the intervention's theoretical mechanism to operate: they are defined not by their surface features but by their functional relationship to the causal pathway the theory of change specifies. Adaptable components are those that serve supporting functions whose specific form can vary across contexts without disrupting the core mechanism, including scheduling, formatting, linguistic register, cultural framing, and resource allocation choices that must respond to local constraints and norms. Distinguishing these categories before evaluation begins enables implementation measurement to focus monitoring attention on fidelity to core components while documenting adaptations to understand how contextual modifications interact with outcomes, rather than treating all implementation variation as deviation to be minimized.

.

The following table provides a structured template for theory of change development that evaluation teams and program developers can use to build the evaluative foundation that learning-oriented evaluation requires.

Table 1. Theory-of-Change Template for Educational Interventions

ToC Element	Guiding Design Question	Example Artifact
Problem statement	What specific, evidence-based problem of practice is the intervention addressing?	Needs assessment documentation; baseline evidence of problem magnitude
Target population	Who is intended to benefit, and which populations may be excluded or differentially affected?	Equity map; subgroup access and participation analysis
Core components	Which elements must be implemented with fidelity for the theory of change to function?	Component map distinguishing core from adaptable elements; fidelity criteria
Causal mechanisms	How do core activities plausibly produce intermediate and long-term outcomes?	Mechanism narrative; identified mediators and their measurement
Implementation conditions	What organizational, professional, and resource conditions enable delivery?	Capacity requirements analysis; leadership routines; support infrastructure
Outcome constructs	What short-term, intermediate, and long-term outcomes does the intervention claim to improve?	Outcome construct definitions; measurement validity justification
Risks and safeguards	What harms to participants or equity could the intervention produce, and how will they be identified and mitigated?	Risk register; privacy protection protocols; equity safeguards

Source: data proceed

The risks and safeguards row of Table 1 represents an evaluative commitment that is underrepresented in standard logic model and theory of change templates but that is essential to equity-sensitive evaluation: every theory of change should include an explicit account of the potential negative consequences, unintended effects, and equity risks that implementation could produce, and should specify the monitoring mechanisms and mitigation strategies through which those risks will be managed. Educational interventions can produce harm through multiple pathways: by consuming instructional time that displaces more effective practices; by imposing workload demands that disproportionately burden already-stretched teachers; by creating assessment or participation demands that disadvantage students with less academic preparation, less digital access, or less familiarity with dominant cultural norms; or by generating data about student performance or behavior that is used in ways that stigmatize individuals or communities. A theory of change that does not anticipate these risks cannot generate the evaluation design needed to monitor them.

Measuring Implementation With Equity-Sensitive Indicators

The second domain translates the theory of change into a structured implementation measurement plan that captures the delivery conditions determining whether the intervention's core mechanism has the opportunity to operate. Implementation measurement in learning-oriented evaluation serves a fundamentally different purpose from compliance monitoring: rather than documenting whether implementation met pre-specified standards for the purpose of accountability attribution, it generates diagnostic information about where implementation is strong, where it is variable, what contextual conditions are associated with higher and lower quality delivery, and which adaptations are improving or undermining fidelity to core components. This diagnostic orientation shapes every aspect of implementation measurement design, from the sampling strategies through which data are collected to the interpretive frameworks through which findings are communicated to implementers.

Equity sensitivity in implementation measurement requires attending to whether the intervention is reaching its intended population with equivalent quality across subgroups that may face differential access barriers, receive differential professional support, or participate in organizational contexts with differential capacity for high-quality delivery. An intervention that is implemented with high average fidelity but with systematically lower quality in schools serving high-poverty communities, or that is reaching only the most academically confident students within classrooms, is producing inequitable implementation conditions that may generate differential outcome effects across subgroups. Identifying these patterns requires implementation data that is disaggregated across the relevant equity dimensions

and interpreted within a framework that attends to the structural factors producing differential implementation quality rather than attributing variation to individual practitioner failure.

Table 2. Indicator Menu for Implementation and Outcomes

Indicator Category	Example Indicators	Interpretation Guardrail	Decision Use
Fidelity to core components	Presence of required routines documented through artifact review; observation fidelity checklist	Use sampling-based checks rather than total surveillance; distinguish core from adaptive variation	Targeted coaching; clarification of non-negotiable routines
Reach and participation	Enrollment and active participation rates disaggregated by relevant subgroups	Interpret low reach in context of structural access barriers; avoid attributing participation gaps to learner motivation	Remove structural barriers; redesign participation supports
Quality of delivery	Observation rubric scores; feedback quality audit results	Use calibrated observation with training; interpret in context of experience and support availability	Improve professional development; refine instructional materials
Learner experience	Perceived clarity of expectations; sense of belonging; experienced workload demands	Collect anonymously where possible; ensure findings lead to visible response actions	Adjust pacing and workload; strengthen community and support routines
Learning outcomes	Performance on construct-valid authentic tasks; validated concept assessments	Use multiple measures; validate constructs against theory of change; avoid weak proxy substitutes	Refine pedagogy; adjust assessment design; revisit theory of change
Equity outcomes	Subgroup performance gaps; accessibility audit results; differential participation patterns	Consent-based data collection; use guardrails against deficit framing; interpret gaps structurally	Equity-targeted support design; task and materials redesign
Cost and feasibility	Teacher time burden; staffing ratios; material and technology costs	Compare costs to alternatives; include hidden costs of preparation and coordination; consider sustainability	Inform scaling decisions; resource allocation planning

Source: data proceed

The interpretation guardrail column is the most conceptually distinctive feature of Table 2 and the feature most directly connected to the learning-oriented evaluation philosophy the framework advances. Each guardrail encodes a specific interpretive principle that should govern how the corresponding indicator is read and communicated, preventing the well-documented tendency of implementation data to be applied in ways that locate responsibility for implementation challenges in individual practitioners or student populations rather than in the systemic conditions that shape both. The guardrail against using reach and participation data to attribute gaps to learner motivation is particularly important in educational contexts where differential participation patterns across demographic groups are frequently explained by reference to deficit characterizations of those groups rather than by analysis of the structural access barriers, cultural mismatch conditions, and institutional support gaps that produce differential participation.

Measuring Outcomes With Valid Constructs and Triangulated Evidence

The third domain addresses the design of outcome measurement that is both methodologically rigorous and substantively aligned with what educational interventions actually claim to improve. The central principle organizing this domain is construct validity: outcome measures should assess the specific constructs that the theory of change identifies as the intervention's intended effects, rather than the constructs that are most easily measurable, most readily available in administrative data systems, or most conventionally used in educational accountability systems. When construct validity is not the organizing principle of outcome measurement selection, evaluations produce findings that are technically sound as measurements of what they measure but evaluatively misleading as assessments of whether an intervention achieved its educational purpose.

Triangulation across multiple outcome measures is both a methodological and an equity principle in learning-oriented evaluation. Methodologically, triangulation compensates for the construct coverage limitations of any single measure: a standardized test and a performance assessment that both claim to measure critical thinking will typically emphasize different aspects of that construct, and their convergence or divergence provides more informative evidence about intervention effects than either alone. As an equity principle, triangulation reduces the risk that evaluation conclusions are driven by measurement artifacts that differentially affect particular student populations: a student population that performs differently on multiple-choice versus open-response assessments may be showing a measurement format effect rather than a construct effect, and triangulation across formats makes that distinction visible.

Treating outcomes as theory-of-change tests rather than as simple dependent variables shifts the evaluative question from "did outcomes improve?" to "did the theory operate as predicted?". This shift requires evaluating whether the mediators specified in the theory of change improved, whether those mediator improvements are associated with outcome improvements in the predicted direction and magnitude, and whether the pattern of findings is consistent with the theory of change mechanism or more consistent with alternative explanations. An intervention designed to improve student writing outcomes by strengthening teacher feedback quality should show, in a theory-of-change evaluation, evidence that feedback quality improved, that improved feedback was associated with more student revision activity, and that revision activity was associated with improved writing outcomes. A pattern in which writing outcomes improved without evidence of improved feedback quality or revision activity should prompt revision of the theory of change, not simply celebration of the outcome result.

Using Evidence to Guide Adaptation and Scaling Decisions

The fourth domain addresses the translation of evaluation evidence into the scaling and adaptation decisions that educational innovation leadership requires. Scaling in education is frequently treated as a deployment decision, a judgment about when an intervention has accumulated sufficient evidence of effectiveness to warrant system-wide implementation. Learning-oriented evaluation reconceptualizes scaling as an ongoing implementation and learning process rather than a one-time threshold decision: the question is not "is this intervention ready to scale?" but "what are the conditions under which this intervention's effects are reliably reproducible, and how can those conditions be created in the next expansion sites while monitoring for the implementation variation and equity impacts that expansion will produce?"

The RE-AIM framework provides a multidimensional structure for scaling readiness assessment that extends beyond effectiveness evidence to address the full range of factors that determine population-level impact. Reach assessment examines whether the intervention has demonstrated the capacity to engage a broadly representative sample of the intended population or whether its positive effects have been demonstrated primarily in self-selected, highly motivated adopter sites that may not represent the characteristics of the full population of sites targeted for expansion. Adoption assessment examines organizational willingness and capacity to take up the intervention across the diversity of settings represented in the intended expansion population. Implementation assessment examines whether delivery quality can be maintained at the fidelity levels associated with positive outcomes as scaling moves beyond the enthusiastic early adopters into the broader population of implementers with more variable motivation, capacity, and organizational support. Maintenance assessment examines whether the conditions that supported initial implementation quality can be sustained beyond the period of external facilitation, funding, and attention that typically accompanies pilot and early-scale phases.

Table 3. Scaling Readiness Checklist for Educational Innovations

Readiness Domain	Key Decision Question	Minimum Evidence Standard
Theory clarity	Are core components and causal mechanisms documented specifically enough to guide consistent implementation across diverse sites?	Documented theory of change; component map distinguishing core from adaptable elements
Implementation feasibility	Can typical sites, with typically available resources and capacity, deliver core components with adequate fidelity?	Pilot feasibility data including workload analysis and resource requirement documentation
Capacity infrastructure	Are professional development, coaching, and implementation support materials in place at the scale required for expansion?	Training plan; coaching model documentation; implementation toolkit
Measurement	Are valid outcome measures and equity-	Indicator definitions and validity

.

plan	sensitive implementation indicators defined with interpretation guardrails?	justification; data governance protocols
Equity safeguards	Are access barriers, differential reach patterns, and subgroup impact variations being actively monitored with ethical data governance?	Equity access audit; disaggregated monitoring plan with consent-based data protocols
Governance clarity	Are decision rights, accountability structures, and escalation pathways for implementation challenges clearly defined?	Governance charter; decision rights documentation; escalation pathway specification
Sustainability plan	Can the intervention be maintained beyond the initial scaling investment at a cost and organizational demand level that is genuinely sustainable?	Multi-year resourcing plan; maintenance routine documentation; exit from external support strategy

Source: data proceed

The sustainability row of Table 3 addresses a scaling failure mode that the educational reform literature documents with particular consistency: the "implementation dip" that occurs when externally supported pilot programs transition to institutionally sustained implementation and discover that the organizational capacity, leadership attention, and resource investment that produced pilot success were not embedded in institutional routines but depended on the temporary conditions of the pilot phase. Scaling readiness assessment that does not include rigorous analysis of sustainability conditions produces scaling plans that are realistic about replication of early-phase outcomes but unrealistic about maintenance of those outcomes after the conditions that produced them are withdrawn.

Discussion

The most consequential interpretive challenge the framework addresses is understanding why educational evaluation so frequently fails to serve the learning and improvement purposes that justify its existence. The failure modes are well-documented and analytically distinct. Measurement validity failures, in which evaluations assess convenient proxies rather than the constructs that interventions claim to improve, produce findings that are technically defensible as measurements of what they measure but educationally misleading as assessments of what matters. These failures are driven by a combination of technical limitation, time pressure, and institutional incentive: measuring construct-valid outcomes takes longer, costs more, and requires methodological expertise that many evaluation contexts do not possess, while administrative data on attendance, course completion, and standardized test scores are readily available and politically legible even when they align poorly with the specific outcomes an intervention targets.

Implementation measurement failures, in which evaluations attribute outcome variation to intervention effects without examining the implementation variation that may explain much of that outcome variation, produce the most persistently misleading findings in the educational evaluation literature. When an evaluation reports that an intervention produced an average effect size of 0.3 standard deviations without examining whether the sites with high implementation fidelity produced substantially larger effects than sites with low fidelity, it generates an evaluative conclusion that conflates the intervention's potential with its average achieved delivery quality. The policy implication of this conflated conclusion is that the intervention works moderately well everywhere, when the more accurate and more actionable conclusion might be that the intervention works strongly in sites with adequate implementation conditions and minimally or not at all in sites without those conditions, a finding that would direct improvement attention toward the implementation conditions that matter rather than toward the intervention design itself.

Equity evaluation failures, in which evaluations report aggregate outcomes without examining differential effects across student subgroups, systematically obscure the possibility that educational innovations produce their strongest benefits for students who are already most advantaged and their weakest benefits for students with the highest needs. This pattern, familiar from decades of educational technology research and documented across diverse intervention types, can produce evaluative endorsement of interventions that widen educational inequalities while appearing to improve average outcomes. The evaluation framework's insistence on equity-sensitive disaggregation is not a methodological nicety but a fundamental ethical commitment: evaluations that do not ask who benefits and who is left behind are not merely incomplete; they are complicit in the inequities they fail to examine.

Utilization failures, in which technically sound evaluation evidence fails to influence the decisions it was designed to inform, represent perhaps the most discouraging pattern in the evaluation literature. Research on evaluation utilization consistently finds that evidence is most likely to influence decisions when it addresses questions that decision-makers experience as genuinely uncertain, when it is communicated in formats and on timelines that align with decision cycles, when it is generated through processes that build the trust and ownership of the decision-makers it is intended to serve, and when the organizational culture treats evidence as a resource for improvement rather than a weapon for accountability (Patton, 2008). Evaluation designs that prioritize methodological sophistication over decision relevance, that communicate findings through academic publications long after the decisions they inform have been made, or that are experienced by practitioners as external judgments rather than collaborative inquiry consistently produce the shelf-filling, influence-minimal evaluation reports that have given evaluation research its unfortunate reputation in many educational policy communities.

Implementing the framework's learning-oriented evaluation principles requires navigating a political economy that is, in many respects, structured to resist them. Funders who require accountability evidence of impact within grant cycles that are shorter than the timescales on which educational interventions produce measurable learning outcomes create incentives for evaluators to report promising proxy indicators as if they were outcome evidence. Policymakers who need to demonstrate that investments in educational innovation are producing results create incentives for evaluators to highlight positive findings and minimize null or negative results. Program developers whose professional identities and organizational survival are invested in their interventions create conditions in which evaluation becomes advocacy rather than inquiry.

Countering these pressures requires deliberate institutional design: establishing evaluation independence through structural separation of evaluation from program development and advocacy functions; building funder and policymaker education about the timescales and evidence standards that credible educational evaluation requires; creating incentive structures that reward honest, uncertainty-acknowledging evaluation reporting rather than penalizing null findings; and developing the evaluation capacity within educational institutions that makes high-quality, learning-oriented evaluation possible without depending entirely on external evaluators whose interests may not align with institutional improvement needs.

The role of trust in creating the conditions for learning-oriented evaluation deserves particular emphasis. Evaluation that is experienced by practitioners as surveillance generates the defensive responses, reporting distortions, and data manipulation that undermine the quality of the evidence it produces. Evaluation that is designed and communicated as a learning resource, that shares findings with implementers before disseminating them externally, that uses data to direct support rather than to assign blame, and that explicitly acknowledges the limitations and uncertainties of its findings builds the trust that makes honest implementation reporting and genuine engagement with challenging evidence possible. This trust is not a soft precondition for rigorous evaluation; it is a methodological requirement for the data quality that rigorous evaluation demands.

For institutional leaders, the framework's most significant practical implication is the integration of evaluation design into innovation planning from the earliest stages rather than treating evaluation as a project activity that follows implementation. When evaluation is retrofitted to existing programs, it frequently discovers that the data needed to answer the most important evaluative questions was not collected during implementation, that the theory of change was never sufficiently specified to support mediator testing, and that the implementation variation across sites is too poorly documented to support contextually informative analysis of differential effects. Leaders who treat evaluation as a planning requirement rather than an accountability afterthought create the conditions for evidence to genuinely inform the adaptation and scaling decisions that innovation management requires.

For researchers, the framework calls for methodological investment in the development of outcome measures that are simultaneously construct-valid, feasible to administer in naturalistic educational settings, and sensitive to the dimensions of learning that educational innovations typically claim to improve. The measurement gap between what educational interventions aim to develop and what available assessment instruments can credibly measure is a significant constraint on the quality of educational evaluation evidence, and closing that gap requires systematic research effort that is currently underinvested relative to the methodological attention devoted to causal identification strategies. Measurement development research that produces valid, practical, openly accessible assessment tools calibrated to the learning constructs that educational innovations target would represent a high-leverage contribution to the quality of educational evaluation practice.

For funders and policy agencies, the framework implies a significant reorientation of evaluation funding priorities: from funding for summative impact studies that produce definitive verdicts on scaled

programs after widespread adoption, toward funding for the developmental evaluation, implementation research, and theory of change testing that can improve innovations during the design and pilot phases when modification is most feasible and least costly. The cost-effectiveness argument for this reorientation is straightforward: identifying and correcting implementation problems and theory of change misspecifications in a pilot phase serving hundreds of students is substantially less costly than discovering those same problems through summative evaluation of a scaled program serving hundreds of thousands of students after the window for cost-effective correction has closed.

The framework's applicability to educational evaluation in low- and middle-income country contexts, including the sub-Saharan African and Global South settings that are particularly relevant to the University of Ghana context in which this paper is situated, requires explicit consideration. Evaluation in these contexts frequently faces constraints that complicate the implementation of the full framework: limited institutional evaluation capacity, underdeveloped measurement infrastructure, short project timelines driven by external funder requirements, and organizational cultures in which evaluation is primarily experienced as an accountability requirement of international donors rather than as an internal improvement resource.

The framework's minimum evidence standards, specified in each row of Table 3, are designed to provide achievable targets for evaluation practice in resource-constrained contexts without compromising the essential evaluative functions that the framework serves. A theory of change with clearly specified core components and mechanisms, implementation monitoring that attends to fidelity and reach using sampling-based rather than comprehensive measurement approaches, outcome measurement that triangulates across at least two construct-valid sources, and equity analysis that examines disaggregated participation and outcome patterns with appropriate consent and interpretive guardrails represents a meaningful and achievable baseline for learning-oriented evaluation even in contexts where comprehensive measurement infrastructure is unavailable. Building indigenous evaluation capacity through training, mentorship, and institutional development is a complementary investment that addresses the structural conditions limiting evaluation quality in these contexts, and represents a more sustainable approach to improving evaluation practice than dependence on external evaluation expertise that leaves no lasting institutional capacity behind.

E. CONCLUSION

Educational innovations can improve outcomes at scale only when implementation and evaluation are designed together from the outset, with sufficient clarity about causal mechanisms to generate interpretable evidence, sufficient attention to implementation conditions to distinguish design quality from delivery quality, sufficient methodological rigor to produce construct-valid outcome inferences, and sufficient equity sensitivity to ensure that evidence of average effectiveness does not obscure differential impacts across the student populations whose advancement is most urgently needed. The learning-oriented evaluation framework proposed in this paper offers practitioners, researchers, leaders, and funders a structured, evidence-grounded architecture for pursuing that integration: specify the theory of change with enough precision to make mediators testable, measure implementation with equity-sensitive indicators interpreted through guardrails that direct data toward support rather than surveillance, measure outcomes with construct-valid instruments triangulated across multiple evidence sources, and treat scaling as a disciplined, evidence-guided learning process rather than a deployment threshold to be crossed. Evaluation designed according to these principles will not eliminate uncertainty about what works in education, but it will generate the kind of contextually rich, equity-attentive, and decision-relevant evidence that makes educational improvement a genuinely cumulative and responsive enterprise rather than a recurring cycle of enthusiastic adoption, disappointing results, and inconclusive debates about what went wrong.

REFERENCES

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-74.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A. and Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*, 50.

Durlak, J. A. and 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.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M. and Wallace, F. (2005). *Implementation research: A synthesis of the literature*. University of South Florida.

.

Glasgow, R. E., Vogt, T. M. and Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health, 89*(9), 1322-1327.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). SAGE.

Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation, 14*(1), 29-48.

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

Spillane, J. P. (2006). *Distributed leadership*. Jossey-Bass.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press.