International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

IMPLEMENTATION AND SCALE WHITE PAPER

From Pilot to National Scale

An Evidence, Implementation, and Sustainability Roadmap for Middle School Mathematics Intervention in the U.S.

Companion publication on responsible implementation and national replication

Arselyn M. Bustoba

Mathematics Educator and Instructional-Design Researcher

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

Executive Summary

Middle school mathematics is a pivotal bridge between elementary foundations and the algebraic reasoning required for high school, postsecondary study, and many technical careers. Yet the national recovery remains incomplete. The 2024 National Assessment of Educational Progress found that the average eighth-grade mathematics score was not significantly different from 2022 and remained below the pre-pandemic 2019 result. The assessment represented approximately 115,200 eighth-grade students in 5,410 schools, making the absence of a meaningful rebound a system-level warning rather than a local anomaly.[1] The challenge is also unequal. In the most recent NCES school-poverty comparison, eighth-grade students in high-poverty schools averaged 271 in mathematics, compared with 293 for students in low-poverty schools.[2]

A promising intervention concept is therefore insufficient. Schools need a disciplined pathway for determining whether a model can be delivered, whether it produces meaningful improvement, for whom it works, what it costs, and whether new sites can reproduce its results without extraordinary external support. This white paper provides that pathway for the companion framework, Developing and Implementing a Scalable, Data-Informed Mathematics Intervention and Teacher-Training Framework for Underserved Middle Schools in the United States. The roadmap is intentionally conservative about scale: expansion is treated as an earned outcome of evidence, not as an automatic consequence of enthusiasm, grant funding, or early adoption.

The roadmap is organized around the SCALE sequence: Select pilot sites; Clarify core and adaptable components; Assess implementation, outcomes, equity, and cost; Learn through planned replication; and Embed the framework through local ownership and sustainable financing. During a readiness period, participating districts establish governance, define the target population, crosswalk priority competencies, verify intervention time, protect student data, and document existing instructional conditions. A first-year pilot then tests feasibility and

preliminary outcomes through repeated six- to eight-week improvement cycles. Subsequent cohorts introduce greater variation in geography, staffing, and student populations while retaining common measures and nonnegotiable instructional principles.

The measurement strategy rejects adoption counts as evidence of success. It combines student competency growth, algebra readiness, retention of learning, access to grade-level content, teacher practice, implementation fidelity, participation, subgroup equity, feasibility, and cost. Impact and implementation questions are specified together, consistent with federal guidance emphasizing that evaluators need information about the intervention, its context, the comparison condition, and the relationship between implementation and observed effects.[4] Fidelity is measured component by component rather than reduced to a single attendance indicator. Costs are collected while the program is operating and valued through an ingredients approach that captures personnel time, facilities, technology, materials, administration, and donated or reallocated resources.[6]

> **CENTRAL RECOMMENDATION** Do not scale a program merely because it can be distributed. Scale only after multiple sites can implement its core components, demonstrate credible learning gains, protect equitable access, and sustain the work at a transparent cost.

1. Why a Pilot-to-Scale Roadmap Is Necessary

1.1 The problem is national, but implementation is local

National mathematics data establish urgency, but they do not identify a single program that every school should purchase or adopt. The 2024 NAEP results show that eighth-grade mathematics performance remained below 2019, while the 2022 poverty analysis documented a 23-point difference between students in high- and low-poverty schools.[12] Those figures describe a common national problem expressed through highly varied local conditions. A rural middle school may struggle with specialist access, transportation, and broadband. An urban school may face student mobility, crowded schedules, and multiple overlapping initiatives. A suburban district may possess strong technology infrastructure but still lack a coherent intervention sequence or adequate time for collaborative analysis.

1.2 Scale is a scientific and organizational question

Education research distinguishes design and development from efficacy, effectiveness, replication, and scale-up. That progression matters because evidence generated under

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

developer-supported conditions may not transfer automatically to routine practice. Federal replication guidance identifies direct and conceptual replication as different ways to test whether findings hold with new samples, settings, implementation supports, measures, or analytic approaches. It emphasizes that replication helps determine what works, for whom, and under what conditions.[14] Responsible expansion should mirror that logic: first stabilize the intervention, then test it in similar settings, then vary context deliberately, and only afterward claim broader applicability.

The originating framework has an appropriate foundation for this progression. Bustoba and Dela Cruz developed and validated a self-paced Mathematics 10 digital module through competency review, module development, and multidimensional expert evaluation. The study supports the content-development method and the resource's acceptability, while also recommending further research on application and efficacy.[15] The present roadmap honors that boundary. It does not convert validation into an unsupported claim of nationwide impact. Instead, it specifies how a U.S. middle-school adaptation can move from a credible design base to feasibility evidence, preliminary impact evidence, replication, and eventually responsible scale.

2. The SCALE Architecture

The SCALE architecture translates the ambition of national reach into five operational commitments. These commitments are sequential but iterative: later evidence may require a return to site selection, component clarification, or measurement design. The architecture is meant to support local adaptation without allowing the intervention to become unrecognizable across sites.

SCALE commitment	Decision question	Required evidence	Primary product
Select	Are the pilot sites appropriate and ready to	Need, leadership commitment, protected time, staffing, data access,	Readiness profile and site agreement

SCALE commitment	Decision question	Required evidence	Primary product
	learn?	technology, and evaluation consent	
Clarify	What must remain constant, and what may change locally?	Theory of action, component specifications, adaptation rules, and version history	Implementation manual, version 1.0
Assess	Was the model delivered, did outcomes improve, was access equitable, and what did it cost?	Fidelity, outcome, equity, feasibility, and ingredients-based cost data	Pilot evaluation and cost report
Learn	Can new sites reproduce the pattern under varied conditions?	Planned replication, contextual documentation, and cross-site analysis	Replication findings and revised manual
Embed	Can local systems maintain quality with declining external support?	Local facilitator capacity, stable financing, governance, and quality assurance	Sustainability and expansion decision

2.1 Core components and adaptable elements

Scale requires a stable identity. The roadmap defines the framework's core components as: a standards-aligned competency architecture; multiple measures of student need; explicit and appropriately intensive mathematics instruction; regular progress monitoring; response-based adaptation; job-embedded teacher learning; fidelity and equity review; and secure, minimal data use. A site cannot claim to implement the framework if these elements are absent. Each core component must be described in terms of content, quantity, mode, and quality, reflecting federal implementation-research guidance.[4]

Other elements may vary. Districts may use different learning-management systems, approved assessments, schedules, staffing configurations, culturally relevant contexts, language supports, and sequences aligned with state standards. Variation is acceptable when it solves a documented local problem without weakening the mathematical goal or reducing access. Every substantive adaptation should record the reason for change, the responsible decision-maker, the expected effect, the date introduced, and the evidence used to retain or reverse it. Version control allows later evaluators to connect outcomes to the program actually delivered.

Instructional content should draw on established recommendations rather than novelty alone. The What Works Clearinghouse algebra guide recommends using solved problems to analyze reasoning, teaching students to recognize the structure of algebraic representations, and helping students choose among alternative solution strategies.[3] The companion REL Algebra Toolkit demonstrates how those recommendations can be embedded in professional learning communities, data tools, classroom trials, and Plan-Do-Study-Act cycles without replacing the local curriculum.[7] These resources offer both content anchors and a model for supporting implementation.

3. Selecting Pilot Sites

3.1 Selection should balance need, readiness, and learning value

Pilot selection is an analytic task, not a competition for the schools with the highest test scores or the most polished applications. The initial cohort should include schools where the need is meaningful, leadership is willing to protect implementation conditions, and the combination of sites can answer useful questions. Selecting only highly resourced schools may produce an easier launch but weak evidence for underserved contexts. Selecting only schools in acute organizational distress may make it impossible to distinguish flaws in the framework from conditions that would undermine nearly any intervention.

Selection criteria must avoid deficit assumptions. Poverty, language status, disability, or geography should never be treated as evidence that students are less capable. The relevant question is whether students have received equitable access to rigorous instruction, appropriate intervention, qualified educators, usable technology, and learning time. Site review should therefore examine opportunity-to-learn conditions alongside achievement data. It should also identify strengths, including experienced teacher leaders, family partnerships, existing intervention blocks, collaborative planning routines, and locally developed resources that can support the pilot.

3.2 The readiness review

Each candidate site should complete a readiness review before final selection. The review begins with a standards and curriculum scan: adopted materials, pacing, prerequisite expectations, current screeners, benchmark schedules, and access to grade-level content. It then examines operations: intervention periods, group-size limits, staff assignments, substitute coverage, professional learning calendars, student transportation, and competing initiatives. Technology review covers devices, connectivity, accessibility, vendor approval, technical support, and printable alternatives. Data review addresses authorized systems, role-based access, parental notices where applicable, retention schedules, and procedures for de-identifying evaluation files.[12][13]

Readiness is rated by domain rather than reduced to a single score. A school may be ready in leadership and curriculum but require support in data governance or scheduling. The implementation team should classify each domain as ready, ready with conditions, or not yet ready. Conditions must name an owner and completion date. A site that lacks an intervention period, cannot protect teacher collaboration time, or has no authorized way to use student data should not launch until the deficiency is addressed. Delaying launch is a quality decision, not a punishment.

3.3 Baseline documentation

Baseline documentation must precede intervention. Student measures should include recent achievement data, a competency-aligned diagnostic or prerequisite check, attendance, course history, and relevant service information. Teacher and system baselines may include content knowledge, confidence, observation of selected practices, intervention schedules, staffing, existing professional learning, and the resources already devoted to mathematics support. Evaluators should also describe usual practice in comparison schools or periods. Without this contrast, a positive change cannot be interpreted accurately because the alternative may include tutoring, commercial software, or other recovery supports.

Baseline data should be disaggregated cautiously to identify access and outcome patterns, not to stereotype students. Small groups require suppression or aggregation to protect confidentiality. Measures must be accessible and appropriate for English learners and students with disabilities. The Student Privacy Policy Office provides federal resources on FERPA, education applications, security, and data governance; participating districts remain responsible for applying federal, state, and local requirements to their specific systems.[13]

4. Implementation Phases and Quality Gates

The roadmap uses four phases over approximately thirty-six months. Timeframes are illustrative because school calendars, procurement, collective bargaining, evaluation design, and local approval processes vary. Progression depends on evidence at each quality gate rather than the passage of time alone.

Phase	Illustrative period	Principal work	Gate for progression
1. Design and readiness	Months 0–3	Select sites; specify components; crosswalk competencies; approve measures, schedules, governance, and baseline collection	All critical readiness conditions have named owners, resources, and completion evidence
2. Feasibility pilot	Months 4–12	Prepare educators; run repeated intervention cycles; monitor fidelity, usability, equity, preliminary outcomes, and actual resource use	Core components are deliverable with acceptable participation and no serious safety, privacy, or equity failure
3. Replication and effectiveness	Months 13–24	Add schools or grades; reduce developer support; test common outcomes; compare contexts; refine facilitator preparation	New sites reproduce implementation and outcome patterns at feasible cost
4. Responsible expansion	Months 25–36+	Establish regional support; publish materials and adaptation rules; pursue stronger evaluation; monitor sustainability	Evidence supports broader use with transparent limitations and durable local capacity

4.1 Phase One: design and readiness

Phase One converts the broad framework into a testable local model. A design team maps state standards and district pacing to priority competencies in ratios and proportional relationships, rational numbers, expressions and equations, functions, geometry, statistics, and algebraic reasoning. Each competency receives a learning progression, prerequisite map, common misconceptions, sample tasks, and mastery evidence. Content panels review accuracy, cognitive demand, cultural and linguistic accessibility, and alignment. Bustoba's prior use of structured expert validation provides a practical precedent for this review process.[15]

The team then defines each support component. Student intervention specifications include eligibility, group size, session length, frequency, duration, instructional routine, progress-monitoring cadence, and response rules. Teacher support specifications include orientation, module study, collaborative planning, classroom application, observation or co-planning, feedback, and evidence review. Data and technology specifications cover approved tools, data fields, access roles, accessibility, low-bandwidth use, and printable options. A logic model connects resources and activities to near-term, intermediate, and long-term outcomes.

Professional learning must be designed as implementation support. The 2025 Title II, Part A guidance describes qualifying professional development as sustained, intensive, collaborative, job-embedded, data-driven, and classroom-focused rather than a stand-alone workshop.[9] The pilot should therefore establish a yearlong sequence before the first student cycle begins. A recent IES mathematics intervention toolkit for grades three through six illustrates the need to budget explicit educator and facilitator time across multiple professional learning community sessions and online activities; while its grade span is not identical, its resource transparency is instructive for sixth-grade planning.[16]

4.2 Phase Two: feasibility pilot

The feasibility pilot runs through repeated six- to eight-week cycles. At the start of a cycle, teachers review diagnostic evidence and assign students to the least intensive support likely to address the identified need. Intervention supplements rather than replaces grade-level instruction whenever feasible. Tier 2 typically provides scheduled small-group support for students with similar gaps. Students with severe and persistent needs receive individualized intensification through smaller groups, additional time, explicit modeling, increased practice, alternative representations, or motivational supports. The National Center on Intensive Intervention emphasizes that adequate assessment, valid progress monitoring, faithful administration, graphing, goal setting, and data-based decisions are integral to intensive intervention.[8]

Each session should follow a coherent routine: activate prerequisite knowledge, state the mathematical goal, model or analyze a worked example, connect representations, provide guided practice, elicit reasoning, provide independent practice, and conduct a brief mastery check. Self-paced digital materials may extend access and feedback, but they should not operate as unsupervised substitutes for teaching. The 2024 National Educational Technology Plan warns that technology can support personalization and performance information while also deepening inequities across access, design, and use.[12] The pilot therefore requires teacher mediation, accessibility checks, printable alternatives, and transparent review of automated features.

Teacher learning proceeds alongside student intervention. Educators meet in grade-level or cross-site professional learning communities to examine tasks, anticipate misconceptions, rehearse questioning, review student work, and plan adaptations. Between meetings, teachers test a small number of practices and collect evidence. Coaches or mathematics leads provide focused observation, co-planning, or feedback. A meta-analysis of causal studies found positive pooled effects of teacher coaching on instruction and student achievement, while also finding smaller effects in larger effectiveness trials; the scale warning supports early development of internal coaching capacity and narrow, observable coaching goals.[10]

Fidelity review during the pilot is formative as well as evaluative. Teachers receive timely information about whether intended dosage, instructional routines, feedback, and progress monitoring occurred. The purpose is to identify barriers and provide support, not to create a punitive compliance system. Federal implementation guidance recommends measurable indicators for direct and support components, component-level scoring, and prespecified thresholds for adequate fidelity.[4] When thresholds lack prior empirical support, the pilot should treat them as provisional and refine them through evidence.

4.3 Phase Three: replication and effectiveness

Replication begins only after the model is sufficiently stable to be taught to others. The next cohort should introduce planned variation: another district, a rural or urban context not represented in the pilot, a different approved assessment system, additional grade levels, or schools with less mature collaborative structures. Variations are selected to answer explicit questions. Unplanned diversity can overwhelm interpretation; planned diversity builds knowledge about generalizability.

Replication evaluation should preserve common primary outcomes and component definitions while documenting contextual differences. Federal replication guidance encourages researchers to specify how populations, intervention components, implementation supports, measures, settings, and analyses differ from prior studies.[14] Both successful and unsuccessful replications are valuable. A null result may reveal a boundary condition, a support component that cannot yet be removed, or a local alternative that performs equally well. Reporting these findings prevents national expansion from being built on selective success.

4.4 Phase Four: Responsible Expansion

Responsible expansion shifts the center of gravity from a development team to a distributed support system. The program publishes an implementation manual, facilitator guides, competency maps, model units, assessment specifications, adaptation logs, fidelity tools, privacy templates, and cost calculators. Regional agencies or district networks provide onboarding and calibration. A central stewardship group maintains version control, reviews evidence, updates materials, and publishes limitations. Local sites retain authority over curriculum and scheduling within the defined component boundaries.

5. Measuring What Matters

5.1 A balanced measurement framework

A credible evaluation combines student outcomes, teacher practice, implementation, equity, feasibility, and cost. State assessments offer comparability but are too infrequent and broad for rapid improvement. District benchmarks are timely but may not be comparable across sites. Curriculum-based measures are sensitive to progress but can narrow the construct. Student work reveals strategy and reasoning but requires scoring calibration. Attendance, participation, and digital completion describe opportunity, not mastery. The framework therefore triangulates evidence rather than allowing one convenient metric to define success.

Primary student outcomes should be prespecified before analysis. A pilot may designate growth on a competency-aligned assessment and performance on an algebra-readiness measure as primary outcomes, with mastery rate, retention after six to ten weeks, state or district benchmark change, attendance, engagement, and course placement as secondary outcomes. Measures must be technically appropriate for the intended decisions. Progress measures used weekly or biweekly should be brief and sensitive to change; broader outcomes should assess transfer rather than merely repeat practiced items.

Teacher outcomes should focus on demonstrated practice. Possible measures include accuracy in diagnosing misconceptions, quality of mathematical representations, use of worked-example analysis, elicitation of student reasoning, responsiveness to progress data, and preservation of grade-level rigor during intervention. Teacher knowledge and confidence can add context but should not substitute for observation, planning artifacts, or student work. A prior IES study of intensive middle-school mathematics professional development found no evidence of improved teacher knowledge or student achievement, illustrating why training exposure alone cannot be treated as an outcome.[11]

5.2 Fidelity, adaptation, and context

Fidelity measures should distinguish adherence, dosage, quality, participant responsiveness, and support received. For example, a small-group intervention component may be measured through sessions scheduled, sessions delivered, minutes delivered, group size, coverage of the intended routine, and observed instructional quality. Coaching may be measured through contact frequency, focus, feedback timeliness, and teacher application. Professional learning may be measured through participation, completion of classroom trials, quality of collaborative analysis, and follow-through on action plans.

5.3 Equity as a decision domain

Equity is not a final subgroup table added after the main analysis. It is a recurring implementation question: Who is referred, who receives the intended dosage, who loses grade-level instructional time, who can access digital materials, who completes the intervention, who improves, and who advances to rigorous coursework? Data should be examined by locally relevant groups when sample sizes permit safe and meaningful interpretation, including economic disadvantage, race and ethnicity, English-learner status, disability, gender, mobility, and school geography.

Differences require investigation rather than automatic conclusions about students. Teams should examine measure accessibility, attendance, staffing, language demands, scheduling, prior opportunity, group composition, teacher assignment, and the quality of implementation. Corrective actions may include accessible formats, multilingual family communication, revised scheduling, additional transportation, stronger instructional support, different grouping, or protection of grade-level content. A program should not pass a scale gate when average improvement conceals material deterioration in access or outcomes for a priority group.

6. Program Evaluation and the Learning Agenda

6.1 Evaluation questions

The evaluation should answer six questions. First, feasibility: can schools deliver the model within ordinary schedules and staffing? Second, fidelity: were direct and support components implemented at intended levels? Third, impact: did students improve relative to an appropriate baseline or comparison? Fourth, equity: were access, implementation, and outcomes fairly distributed? Fifth, cost: what resources were required, and how did cost vary by site and phase? Sixth, sustainability: can schools maintain quality with reduced external support? These questions should be mapped to measures, data sources, timing, responsible analysts, and decision rules before launch.

The feasibility pilot may use a matched comparison, comparative interrupted time series, regression discontinuity where assignment rules permit, or another credible quasi-experimental design. If operationally and ethically feasible, later studies may use cluster random assignment at the school or grade-team level. The design must match the claim. A pre-post increase among participating students is encouraging but cannot establish that the framework caused the improvement. Results should state limitations, attrition, baseline differences, missing data, measure changes, and the services received by comparison students.

6.2 Evidence progression and independent review

Evidence should progress from usability and feasibility to preliminary associations, then to stronger causal evaluation and replication. ESSA identifies tiers of strong, moderate, and promising evidence based on research design and additional considerations, while also recognizing activities that demonstrate a rationale and include ongoing evaluation.[5] The framework should label its current evidence status accurately. A positive first-year pilot may justify refinement and replication; it should not be described as nationally proven.

Independent evaluation strengthens credibility. The development team should participate in defining the program and interpreting implementation, but an external evaluator or methodologist should prespecify primary outcomes, oversee analysis, and report unfavorable as well as favorable findings. Replication guidance recommends safeguards for objectivity when original investigators remain involved.[14] Preregistration, documented analysis plans, measure dictionaries, version histories, and reproducible code should be used when permitted by privacy rules and available resources.

6.3 Decision rules

Before data are examined, the governance group should define outcomes that support continuation, revision, pause, or termination. Continuation may require acceptable implementation of all critical components, positive movement on the primary outcome, stable or improved subgroup access, and manageable burden. Revision may be appropriate when the theory remains plausible but a component is confusing, under-supported, or too costly. A pause is warranted for data-governance failures, inaccessible materials, chronic scheduling failure, or inability to collect valid outcomes. Termination should remain possible when repeated, adequately implemented trials do not produce meaningful benefit or when cost and burden are disproportionate.

> **EVIDENCE RULE** A scale decision must explain not only whether outcomes changed, but also what was delivered, who received it, what alternatives existed, what it cost, and whether another site reproduced the result.

7. Cost Effectiveness and Financial Sustainability

7.1 Measure resources from the beginning

Cost is part of implementation, not an appendix written after the pilot. The IES Cost Analysis Starter Kit recommends clarifying the program model, choosing the perspective, identifying and pricing ingredients, adjusting for context, calculating costs, testing sensitivity, and determining whether further economic evaluation is appropriate.[6] The roadmap uses this ingredients method from the first month so that schools record actual resources while memories and records are current.

Key ingredients include instructional coordinator time, teacher planning and delivery time, facilitator preparation, coaching, substitute coverage or stipends, assessment administration and scoring, module development, accessibility review, technology licenses, devices attributable to the program, printing, facilities, data management, administration, evaluation, and family engagement. Existing resources are not costless. A teacher's planning period, a district analyst's time, or donated university support has an opportunity cost even when no new invoice is paid. Reports should distinguish total economic cost from new budget expenditure and incremental cost above usual practice.[6]

7.2 Report cost in decision-ready forms

District leaders need more than one total. The evaluation should report cost per participating student, cost per school, cost per teacher trained, cost per completed intervention cycle, and, when outcome evidence permits, cost per additional student reaching the defined competency threshold. Costs should be separated into start-up and recurring categories. Start-up costs may include initial design, training, data integration, and facilitator preparation. Recurring costs include intervention delivery, progress monitoring, coaching, platform use, maintenance, and quality assurance.

7.3 Financing strategy

Sustainability requires alignment with existing systems. Whenever appropriate, the framework should use adopted curricula, approved platforms, established intervention blocks, common planning time, district data systems, and teacher-leader roles. Potential funding may include state and local instructional funds, school-improvement resources, special education or Title I resources when the activity and students meet applicable requirements, and Title II, Part A support for qualifying professional learning. The 2025 Title II guidance recognizes sustained, intensive, collaborative, job-embedded, data-driven, and classroom-focused professional development and discusses activities such as professional learning communities, data use, mentoring, and STEM learning, subject to statutory and local requirements.[9]

A braided financing plan must preserve compliance and program coherence. Districts should document which fund supports which ingredient, the applicable allowability determination, supplement-not-supplant considerations where relevant, and the plan when temporary grants end. Expansion should not rely on unpaid labor or heroic effort. If the model cannot operate without chronic after-hours work, donated specialist time, or intensive developer intervention, it has not yet demonstrated sustainability.

8. Partnerships and Governance

No single organization should own every function. Districts and schools own curriculum alignment, staffing, scheduling, student support, family communication, privacy compliance, and day-to-day implementation. The framework's stewardship team maintains component definitions, version control, training standards, common measures, and cross-site learning. Universities or independent evaluators support study design, analysis, and dissemination. Regional educational agencies can convene districts, prepare facilitators, and reduce duplication. Mathematics associations and content experts can review accuracy and rigor. Technology partners must meet district accessibility, security, interoperability, and data-use requirements.

Governance should operate at three levels. A school implementation team reviews weekly operations and individual student response. A district team reviews staffing, schedules, professional learning, fidelity, equity, and cross-school patterns at least monthly during active cycles. A cross-site council reviews evidence, adaptations, cost, and scale decisions each semester. Membership should include mathematics teachers, intervention or special education expertise, English-learner representation, school leadership, data and technology staff, and, where appropriate, students and families.

9. Responsible National Expansion

9.1 A cohort model for national reach

A national learning network should publish annual evidence summaries that include setbacks, adaptations, contextual differences, and unresolved questions. Materials should carry version numbers and evidence labels. Sites should know whether a recommendation is supported by causal evidence, descriptive implementation evidence, expert validation, or a locally promising practice. The network should avoid product-centered claims and remain compatible with varied state standards and district curricula.

9.2 Expansion gates

Five gates protect responsible growth. The readiness gate asks whether sites possess the conditions to begin. The implementation gate asks whether critical components can be delivered with acceptable quality. The outcome gate asks whether students show meaningful improvement on prespecified measures. The equity gate asks whether access and benefits extend to the intended populations without removing rigorous instructional opportunity. The sustainability gate asks whether cost, staffing, governance, and local expertise can support continued operation. Failure at one gate triggers corrective action or delay, not concealment.

The strongest signal for scale is not a large first-year effect. It is a coherent pattern across sites: the model is implementable, outcomes move in the expected direction, priority groups receive the intended support, costs become more efficient without loss of quality, and new sites succeed with less external assistance. Replication is especially important because school contexts vary and because findings from one setting may depend on unusually skilled leaders or favorable conditions.[14]

9.3 Knowledge products

The initiative should produce public knowledge as well as services. Priority products include a pilot-site readiness tool, competency-crosswalk protocol, intervention-unit template, facilitator curriculum, fidelity rubric, adaptation log, equity review protocol, cost workbook, data dictionary, evaluation template, and annual learning report. These materials allow districts to understand what implementation requires before adoption and support independent scrutiny of claims.

10. Risks and Safeguards

Risk 1 — Premature expansion. Demand may grow faster than evidence. The safeguard is a cohort cap and public progression gates that require replication, equity review, and feasible cost.

Risk 2 — Fidelity becomes compliance. Checklists can encourage superficial delivery. The safeguard is component-level measurement combined with teacher feedback, student work, observation, and problem-solving support.

Risk 3 — Local adaptation fragments the model. Untracked changes can make outcomes uninterpretable. The safeguard is a core-versus-adaptable specification, approval rules, adaptation logs, and version control.

Risk 4 — Technology substitutes for instruction. Digital completion may be mistaken for learning. The safeguard is teacher-mediated use, mathematical reasoning tasks, accessible and printable alternatives, and outcome measures beyond platform data.[12]

Risk 5 — Average gains conceal inequity. A program may improve overall while priority groups receive less dosage or lose grade-level content. The safeguard is recurring disaggregation of access, implementation, growth, and placement.

Risk 6 — Hidden labor undermines sustainability. Programs may appear inexpensive because teacher, facilitator, or partner time is unpriced. The safeguard is ingredients-based cost analysis that values reallocated and donated resources.[6]

Risk 7 — Evaluation serves promotion. Selective outcomes and favorable sites can exaggerate success. The safeguard is prespecified outcomes, independent review, transparent limitations, and publication of null or adverse findings.

Risk 8 — Data collection exceeds instructional need. Collecting more information increases burden and privacy exposure. The safeguard is data minimization, role-based access, retention schedules, district approval, and de-identified public reporting.[13]

Conclusion

The United States does not need another mathematics intervention that reaches many schools before anyone knows whether it can be implemented well. It needs a learning system that moves carefully from validated design to feasible delivery, credible evaluation, planned replication, sustainable financing, and responsible expansion. The SCALE architecture provides that pathway: select sites for need and learning value, clarify the model, assess outcomes and implementation together, learn through replication, and embed the work in local capacity.

This roadmap treats underserved schools as partners in development rather than passive recipients of a finished product. Teachers test and improve routines; students and families identify barriers; districts align the work to standards and operations; evaluators protect the integrity of evidence; and regional partners extend learning across jurisdictions. Cost, equity, privacy, and implementation quality are measured from the beginning, not addressed after expansion.

National scale is ultimately a claim about reproducibility and durability. The framework should expand only when diverse schools can deliver its core components, students demonstrate meaningful progress, opportunity gaps do not widen, local educators can sustain the work, and transparent cost information supports adoption. By making evidence the condition for growth, the initiative can build something more valuable than a widely distributed program: a credible, adaptable, and enduring system for strengthening middle school mathematics instruction.

Source Notes

1. **U.S. Department of Education, National Center for Education Statistics.** 2024 NAEP Mathematics Assessment: Results at Grades 4 and 8 for the Nation, States, and Districts, NCES 2024-217.

2. **National Center for Education Statistics.** Condition of Education: Mathematics Performance.

3. **What Works Clearinghouse.** Teaching Strategies for Improving Algebra Knowledge in Middle and High School Students.

4. **Institute of Education Sciences.** Conducting Implementation Research in Impact Studies of Education Interventions: A Guide for Researchers.

5. **What Works Clearinghouse.** ESSA Tiers of Evidence: What You Need to Know.

6. **Institute of Education Sciences.** Cost Analysis: A Starter Kit, Version 1.1.

7. **Regional Educational Laboratory Central.** Toolkit to Support Evidence-Based Algebra Instruction in Middle and High School.

8. **National Center on Intensive Intervention.** Mathematics Progress Monitoring Training Modules.

9. **U.S. Department of Education.** Title II, Part A of the Elementary and Secondary Education Act of 1965: Non-Regulatory Guidance, January 2025.

10. **Kraft, M. A., Blazar, D., and Hogan, D..** The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence.

11. **Institute of Education Sciences.** Middle School Mathematics Professional Development Impact Study.

12. **U.S. Department of Education, Office of Educational Technology.** A Call to Action for Closing the Digital Access, Design, and Use Divides: 2024 National Educational Technology Plan.

13. **U.S. Department of Education, Student Privacy Policy Office.** Protecting Student Privacy.

14. **Institute of Education Sciences and National Science Foundation.** Companion Guidelines on Replication and Reproducibility in Education Research.

15. **Bustoba, A. M., and Dela Cruz, M. C. A..** Development and Validation of Self-Paced Learning Digital Module in Mathematics 10.

16. **Regional Educational Laboratory Northeast & Islands.** Mathematics Intervention Toolkit: Resources for Toolkit Success, REL 2026-004.

Author Evidence Base

This publication is a companion to the white paper Developing and Implementing a Scalable, Data-Informed Mathematics Intervention and Teacher-Training Framework for Underserved Middle Schools in the United States. Its design foundation also draws on Bustoba and Dela Cruz's published module-development study, cited in Source Note 15, and on the professional résumé of Arselyn M. Bustoba supplied in July 2026.

About the Author

Arselyn M. Bustoba is a mathematics educator and instructional-design researcher with experience teaching middle and secondary grades in the United States and the Philippines. She holds graduate preparation in mathematics education and has professional experience in curriculum development, formative assessment, differentiated and individualized intervention, digital learning, and teacher mentoring. Her work focuses on building practical systems that connect rigorous mathematics content, timely evidence, and educator capacity.