All Learners Network:

Evaluation of Impact on Mathematics Teaching and Student Achievement

Authored by John Tapper, PhD
Founder and Chief Executive Officer, All Learners Network
December 2025

Executive Summary

This evaluation synthesizes findings from impact studies conducted across ten school districts and supervisory unions that partnered with All Learners Network (ALN) between 2017 and 2025. The analysis examines both quantitative student achievement data and qualitative evidence of instructional change across diverse educational contexts including rural Vermont supervisory unions, suburban Massachusetts districts, high-poverty Maryland schools, mid-sized New Hampshire districts, and a large Colorado district serving over 15,000 students. ALN worked with teachers in preK-12.

Across examined sites, student mathematics achievement improved following ALN implementation, with gains most pronounced in sites that implemented embedded coaching models alongside professional development workshops. Effect sizes ranged from moderate (partial eta squared = 0.32-0.47 in some middle school cohorts) to large (partial eta squared = 0.74-0.91 in most elementary settings). The typical pattern of improvement indicates instructional changes occurring in Year 1 of implementation, with measurable student achievement gains typically emerging in Year 2.

Notably, achievement gains were consistently larger among historically underserved student populations. Students receiving special education services, students from economically disadvantaged backgrounds, and English language learners demonstrated accelerated growth compared to general education peers in multiple sites. This pattern was consistent across different assessment types (state assessments, iReady diagnostic, and VTCAP) and demographic contexts.

The evidence base has limitations. No studies employed randomized control designs or included matched comparison groups. Implementation fidelity data varies across sites. The amount of time allocated to professional learning, the administrative support, and the rate of teachers adoption were all variables for fidelity. Several promising cases lack robust quantitative outcome measures, while others provide test score data without detailed implementation information. These limitations constrain causal inference while providing suggestive evidence of positive association between ALN implementation and improved student outcomes.

Introduction: The ALN Model and Theory of Change

All Learners Network's professional development model operates from a theory of change centered on three interconnected instructional practices:

High Leverage Concepts (HLCs): ALN focuses teacher attention on conceptual anchors that support learning across grade levels. Core concepts include early number sense, additive reasoning, multiplicative reasoning, and proportional relationships. This focused approach theoretically allows for deeper conceptual development.

Differentiated Instruction through Main Lesson and Math Menu: The ALN model structures mathematics instruction into two components: Main Lesson and Math Menu. During Main Lesson, all students access grade-level content through differentiated entry points matched to their conceptual understanding. During Math Menu, teachers provide intensive small-group instruction based on formative assessments while other students engage in independent practice. Instruction during Math Menu is focused on “just right” learning for each student, challenging some students and supporting unfinished learning with others. This structure enables both inclusive grade-level participation and targeted intervention.

Formative Assessment as Instructional Practice: ALN trains teachers in assessment protocols including work sorts that analyze student mathematical thinking to inform instructional decisions. Assessment becomes a tool for understanding student reasoning patterns rather than solely measuring accuracy.

The theory of change predicts that implementing these practices will result in: increased engagement among students who previously struggled or disengaged; narrowing achievement gaps between high and low performers; and accelerated growth for students receiving special education services as they receive instruction matched to conceptual understanding rather than grade-level placement.

Quantitative Findings Across Sites

Thompson School District, Colorado (2022-2025)

Context: Thompson School District serves over 15,000 students across 23 elementary and middle schools in Northern Colorado. The district implemented ALN professional development district-wide beginning in 2022-23, with all mathematics teachers participating in workshop sessions and many receiving embedded coaching support.

Methodology: Repeated measures ANOVA examined student performance on iReady mathematics diagnostic assessments across three consecutive winter testing windows (2022-23, 2023-24, 2024-25). The analysis tracked individual students over time, controlling for baseline performance. Sample sizes ranged from approximately 150 students per cohort at smaller schools to over 500 students at larger schools.

Key Findings:

- Every school (n=23) demonstrated statistically significant improvement in mean mathematics scores over the three-year period ($p < .001$ for 22 schools; $p < .05$ for one school with small sample)
- Effect sizes were large at most elementary schools, with partial eta squared values ranging from 0.66 to 0.91, indicating substantial proportions of variance in achievement gains attributable to time/intervention
- Middle schools showed significant but more modest effect sizes (partial eta squared 0.19-0.70), suggesting differential impact by grade configuration (elementary vs middle school)
- Pairwise comparisons confirmed consistent year-over-year growth, with no schools showing decline during implementation years
- Mean scale score improvements ranged from 14.9 points (Walt Clark Middle) to 58.0 points (Truscott Elementary) over the three-year period

The consistency of improvement across 23 diverse schools (serving populations ranging from affluent suburban to high-poverty contexts) strengthens confidence in the generalizability of findings within similar districts.

Winooski Public Schools, Vermont (2023-2025)

Context: Winooski is Vermont's most diverse school district, with 44% English language learners and 68% economically disadvantaged students. The district serves approximately 750 students K-12.

Methodology: Repeated measures analysis of Vermont Comprehensive Assessment Program (VTCAP) data across three years. The analysis tracked individual student cohorts as they progressed through grade levels.

Key Findings:

- Grade 7 cohort (students who began ALN-trained instruction in Grade 5) showed large, statistically significant growth ($p < .001$, effect size 0.619), with mean scale scores increasing from 1,654 (2023) to 1,714 (2025)
- Grade 6 cohort showed substantial growth ($p < .001$, effect size 0.545), with most dramatic improvement occurring between Year 2 and Year 3
- Grade 4 cohort showed no significant change, suggesting differential effectiveness based on implementation conditions or grade-level factors
- English language learners demonstrated accelerated growth compared to native English speakers in responding cohorts

The variable outcomes across grade cohorts within a single district illustrate the importance of implementation conditions. Qualitative data from teacher surveys suggest differential adoption rates and varying coaching support across grade levels.

Keene School District, New Hampshire (2022-2024)

Context: Mid-sized district serving approximately 2,800 students. Keene invested in the initial year (2022-23) by engaging in needs assessment and coach development before implementing classroom-level professional development in 2023-24.

Methodology: Pre-post analysis using New Hampshire SAS (Statewide Assessment System) mathematics scores, disaggregated by student subgroups.

Key Findings:

- District-wide proficiency increased 24% in first implementation year (following planning year)
- Students receiving special education services improved 16%
- Students from economically disadvantaged backgrounds showed 35% increase in proficiency

The rapid first-year gains following an infrastructure-building year suggest potential value in delayed implementation models, though absence of comparison conditions precludes definitive conclusions about optimal implementation.

Worcester County Public Schools, Maryland (2017-2023)

Context: Five high-poverty elementary schools where PARCC mathematics scores substantially trailed state averages.

Methodology: Longitudinal tracking of state assessment proficiency rates from 2017 baseline through 2023, six years post-intervention.

Key Findings:

- One school (Pocomoke Elementary) increased proficiency 30 percentage points in first implementation year
- In 2023, six years after initial intervention, district mathematics performance exceeded state averages by 23 percentage points
- Sustained outperformance for both students in poverty and students with disabilities relative to state benchmarks

The six-year sustained improvement addresses critical questions about professional development durability. However, absence of detailed implementation fidelity data between 2018-2023 limits understanding of maintenance mechanisms.

Additional Implementation Sites

Plymouth Public Schools, Massachusetts: Five-year partnership with focus on High Leverage Concepts and CRA (Concrete-Representational-Abstract) assessment protocols. Student engagement with mathematics instruction increased measurably. Teachers reported significant shifts in formative assessment practice and differentiation strategies. Quantitative achievement data unavailable for this analysis.

C.P. Smith Elementary, Vermont: Students receiving special education services achieved proficiency rates three times state averages for comparable populations on VTCAP assessments. Small school context (n~200 students) limits generalizability.

Orange East Supervisory Union, Vermont: Statistically significant improvement in state assessment outcomes following two-year implementation. Effect sizes ranged from moderate to large across participating schools.

Qualitative Findings on Instructional Change

Multiple sites collected teacher survey and interview data documenting self-reported changes in instructional practice. While self-report data has inherent limitations, patterns emerged consistently across diverse contexts. The convergence of teacher reports across different geographic regions, school contexts, and implementation timelines strengthens confidence in findings despite the absence of systematic classroom observation data at most sites.

Differentiation and Grouping Practices

Teachers across multiple sites reported fundamental shifts in how they structure mathematics instruction time. The Math Menu framework was most frequently cited as changing practice. Teachers described moving from whole-group instruction toward flexible small-group models, with grouping based on formative assessment rather than fixed ability tracking.

Evidence from Plymouth Public Schools: End-of-year teacher surveys (n=47 teachers across 5 years of implementation) asked respondents to describe their instructional time structure before and after ALN professional development. In baseline surveys (collected retrospectively), 89% of teachers reported using whole-group instruction for entire mathematics periods, with differentiation occurring primarily through varied worksheet assignments or "challenge problems" for early finishers.

After two years of implementation, at the middle and high school levels, 73% of surveyed teachers reported implementing Math Menu structures at least three times weekly. Teachers described typical structures as:

- 20-25 minutes Main Lesson with all students accessing grade-level content through differentiated entry points
- 25-30 minutes Math Menu with teacher working intensively with 6-8 students while others engage in independent practice, games, or technology-based activities
- 10 minutes reflection/sharing where students explain mathematical thinking

Notably, 68% of teachers reported that their Math Menu groups changed weekly or bi-weekly based on formative assessment results, contrasting with previous practices where students remained in fixed ability groups throughout the year. One third-grade teacher's survey response exemplifies the shift: "I used to have my 'low,' 'medium,' and 'high' groups and they basically stayed there all year. Now I'm constantly regrouping based on what the work sorts show me about their thinking.

A student might be in my intensive group for place value but working independently on geometry."

Evidence from Worcester County: Teacher interview data (n=23 teachers, collected in Year 2 of implementation) revealed substantial changes in how special education teachers participated in mathematics instruction. Before ALN implementation, special education teachers at all five participating elementary schools operated pull-out programs where students with IEPs received all their mathematics instruction in separate settings, typically focused on procedural practice of computation skills.

By Year 2, four of five schools had restructured special education mathematics support as push-in collaborative teaching. Special education teachers described their new role as co-facilitating Math Menu, which allowed them to work with flexible small groups including but not limited to students with IEPs. As one special education teacher explained in interviews: "Instead of taking my kids out to drill addition facts, I'm in the classroom during Math Menu. Sometimes I'm working with students with IEPs on place value concepts, but sometimes I'm working with students who just need more time on that concept, whether they have an IEP or not. And I get to see what grade-level work looks like, so I can build toward that instead of just practicing procedures in isolation."

This shift had measurable implications. Student schedule analysis showed that students with IEPs who previously received 30-45 minutes daily of pull-out mathematics intervention plus 45 minutes of grade-level mathematics (total 75-90 minutes, but segregated) now received 60-75 minutes of inclusive mathematics instruction with differentiated support structures. Administrator interview data indicated that this change initially faced resistance from some special education staff concerned about students "falling further behind," but student achievement data showing accelerated growth for students with IEPs (culminating in C.P. Smith's finding of proficiency rates three times state averages) reduced resistance over time. Evidence from Thompson School District: Instructional coach reports and administrator interviews documented implementation patterns. District mathematics coaches conducted monthly observation visits to all 23 schools using a structured observation protocol that included tallying instructional time allocation.

Observation data from Year 1 (fall observations, n=115 classroom visits) showed:

- 67% of classrooms used whole-group instruction for majority of mathematics period
- 28% implemented Math Menu structures (though with variable quality as noted in coach field notes)
- 5% used other small-group models not aligned with ALN framework

By Year 2 (spring observations, n=118 classroom visits):

- 31% of classrooms used whole-group instruction for majority of mathematics period
- 64% implemented Math Menu structures (with coach notes indicating improved quality: "more purposeful grouping," "tighter alignment between formative assessment and group composition")
- 5% used other models

The shift from 28% to 64% implementation of Math Menu structures between Year 1 and Year 2 aligns with the pattern of instructional change preceding achievement gains, that is expanded upon below. The observation data also revealed variability, with some schools showing near-universal adoption (e.g., Ivy Stockwell Elementary with 15 of 16 teachers implementing Math Menu by Year 2) while others showed more scattered adoption (e.g., Walt Clark Middle School with 8 of 17 teachers implementing by Year 2)—patterns that correlate with the differential effect sizes observed across schools.

Evidence from Winooski: Teacher focus group data (conducted at end of Year 2, n=14 teachers) provided insight into implementation challenges and successes. Teachers reported that the Math Menu structure was particularly valuable in Winooski's context given the wide range of mathematical preparation levels among students, many of whom are recent immigrants with interrupted formal education.

One fifth-grade teacher's focus group comments illustrate the shift: "Before, I had students who couldn't yet do multi-digit addition sitting through whole-group lessons on fraction multiplication. They'd shut down. Now during Main Lesson, I can give everyone access to the fraction concept through different entry points—some using fraction tiles, some drawing pictures, some moving toward abstract. Then during Math Menu, I can pull the group that needs more work on place value or basic multiplication facts that are foundational for fraction work. They're not missing the grade-level concept, but they're getting the foundational support they need."

However, focus group data also revealed implementation challenges. Grade 4 teachers (the cohort showing no significant achievement gains) reported difficulty implementing Math Menu given classroom size (28-32 students), limited space for centers, and pressure to maintain curriculum pacing for benchmark assessments. This qualitative data helps explain the quantitative finding of differential outcomes across grade levels within the same district.

Formative Assessment Use

Multiple sites documented increased use of formative assessment protocols, particularly work sorts where teachers analyze student work samples to identify reasoning patterns. Teachers reported this practice shift occurring earlier than changes in student achievement data, consistent with the theory of change predicting instructional changes preceding achievement gains.

Evidence from Orange East Supervisory Union: Teacher survey data tracked formative assessment practice adoption across two implementation years. Baseline surveys (administered at start of Year 1, n=31 teachers) asked teachers to estimate frequency of various assessment practices:

Baseline (pre-implementation) self-reported practices:

- Analyzing student work to identify patterns in mathematical thinking: 19% reported weekly or more frequently
- Using assessment results to form instructional groups: 32% reported weekly or more frequently
- Conducting structured work sorts: 0% (practice was unfamiliar to all surveyed teachers)

Year 1 end-of-year survey (n=29 teachers, 2 teachers left district):

- Analyzing student work to identify patterns: 72% reported weekly or more frequently
- Using assessment results to form instructional groups: 79% reported weekly or more frequently
- Conducting structured work sorts: 66% reported weekly or more frequently

Notably, student achievement data did not show statistically significant gains until Year 2, despite these substantial self-reported practice changes in Year 1. The pattern of changes in teaching practices preceding achievement change appeared consistently across sites and supports ALN's theory that instructional shifts require time to translate into measurable student outcomes.

Evidence from Plymouth Public Schools: In addition to teacher surveys, Plymouth conducted limited classroom observation using a structured protocol during Years 2-3 of implementation (n=62 observations across 28 teachers). Observers documented specific formative assessment practices during 60-minute mathematics instruction periods.

Observation data revealed:

- 81% of observed lessons included teachers collecting student work for analysis (either during class or for later review)
- 58% of observations included teachers examining student work with colleagues during planning time (observed during pre- or post-lesson planning periods)
- 43% of observations captured teachers explicitly adjusting instruction during the lesson based on student responses (e.g., extending time on a concept, providing additional scaffolding, or advancing more quickly than planned)

Observer field notes provided qualitative detail. One observer noted: "Teacher collected exit tickets showing student attempts at comparing fractions. During Math Menu the next day, had clearly used exit tickets to form groups—one group all showed 'counting pieces' strategy (treating 2/3 as '2 and 3'), another group used visual models accurately but struggled with notation, third group ready for more complex comparisons. Teacher's instruction to each group was clearly differentiated based on yesterday's formative assessment."

The observation data provided validity evidence for teacher self-reports while also revealing variability. Teachers who self-reported high frequency of formative assessment use showed corresponding behaviors in observations. However, some teachers self-reported practices that were not evident in observations, suggesting either social desirability effects in survey responses or variable implementation across different lesson topics or days.

Evidence from Keene School District: Teacher interview data (n=18 teachers, semi-structured interviews conducted at end of Year 1 of classroom implementation) revealed teachers' conceptual understanding of formative assessment purposes had shifted. Researchers coded interview transcripts for teachers' articulated purposes of assessment.

When asked "What is the purpose of assessment in mathematics?", representative responses included:

Before ALN implementation (baseline interviews, reconstructed from teachers' descriptions of previous practice):

- "To see if students got the right answer" (n=11 teachers)
- "To assign grades" (n=14 teachers)
- "To determine if students are ready to move on" (n=8 teachers)

After Year 1 of implementation:

- To understand how students are thinking about mathematical concepts" (n=16 teachers)
- "To figure out what students understand so I can plan instruction" (n=15 teachers)
- "To identify patterns in student reasoning that tell me what they're ready for next" (n=12 teachers)

Several teachers described fundamental reconceptualization of the purpose of assessment, from measuring correctness to understanding mathematical thinking. One fourth-grade teacher explained: "I used to look at a worksheet and just count how many they got right. Now when I do a work sort, I'm looking at how they got their answers. Did they use a visual model? Did they rely on memorized procedures? Are they showing multiplicative thinking or are they still additive? That tells me so much more about what they actually understand and what they need next."

Another teacher articulated the connection between formative assessment and differentiation: "The work sorts changed everything for me. I can't do Math Menu without them because I wouldn't know how to group students. And once I started really looking at student thinking, I realized that kids I thought 'got it' because they had right answers were sometimes using fragile procedures, and kids I thought didn't understand sometimes had really solid conceptual thinking but made computational errors. You can't see that unless you look at the work carefully."

The interview data suggests that teacher understanding of formative assessment purposes may be as important as frequency of practice. Teachers who articulated formative assessment as a tool for understanding thinking (rather than measuring correctness) showed stronger implementation fidelity on observation protocols conducted during Year 2.

Conceptual Focus versus Coverage

Teacher interview data suggested tension between ALN's High Leverage Concepts approach and district pacing guides or curriculum mandates. Teachers who successfully navigated this tension reported feeling supported by building administrators to prioritize depth over coverage. Districts with stronger qualitative and quantitative outcomes generally demonstrated administrative support for instructional model implementation.

Evidence from Orange East Supervisory Union: Teacher interviews (n=24, conducted midway through Year 2) specifically probed tensions between High

Leverage Concepts focus and existing curriculum structures. Of 24 interviewed teachers, 21 (88%) reported feeling tension between ALN's recommendation to spend extended time on foundational concepts versus district-adopted curriculum pacing guides.

One third-grade teacher described the dilemma: "Our math curriculum has us teaching multiplication in six lessons over two weeks, then we're supposed to move to division. But my students weren't solid on equal groups and arrays—they were still counting by ones. ALN training [suggested] to slow down and really develop the multiplicative reasoning. But my principal was asking why we were behind the pacing guide."

However, outcomes varied based on administrative response. Teachers reported two distinct patterns:

Pattern A - Administrative support for depth (n=14 teachers): These teachers reported that after discussing the conceptual focus with administrators, they received explicit permission to deviate from pacing guides when formative assessment indicated students needed more time on foundational concepts. One teacher explained: "My principal said, 'If the work sorts are showing they don't have place value, stay there as long as needed. We'll deal with the pacing guide later.' That permission changed everything. I spent four weeks on place value when the curriculum said two, and those students went on to be much more successful with later content that depends on place value understanding."

Pattern B - Pressure to maintain pacing (n=10 teachers): These teachers reported continued pressure to maintain curriculum pacing despite student need for additional time on concepts. One teacher described: "I know my students needed more time on fractions, but we had a benchmark assessment coming up that covers the whole curriculum. I had to keep moving even though I knew they weren't ready. It felt like we were back to teaching procedures without understanding."

Analysis of Year 2 student achievement data by teacher revealed a correlation: teachers reporting Pattern A (administrative support for depth) had students showing average gains of 27 scale score points on state assessments, while teachers reporting Pattern B (pressure to maintain pacing) had students showing average gains of 11 scale score points. While this analysis cannot establish causation given the small sample and lack of control for other variables, the differential outcomes are consistent with the hypothesis that administrative support for the instructional model affects implementation quality and student outcomes.

Evidence from Winooski: The differential outcomes across grade levels within Winooski (Grade 7 showing large gains, Grade 4 showing no significant change) correlate with qualitative data about administrative support and curriculum alignment challenges. Administrator interviews revealed that Grades 5-7 had building principal support for adjusting pacing, while Grade 4 faced pressure to maintain pacing for district benchmark assessments.

Grade 4 teacher focus group data: "We have benchmark assessments every six weeks that are directly from our curriculum. If we slow down to really develop a High Leverage Concept, our students bomb the benchmarks because we haven't 'covered' all the topics. Then we're in meetings explaining why our scores are low. It's easier to just follow the curriculum even if we don't think students are really learning."

In contrast, Grade 5-7 teacher focus group data: "Our principal told us at the start that she was more interested in seeing students develop real mathematical thinking than checking off topics. She said the state test is about concepts, not specific lessons from our curriculum, so if we focused on the High Leverage Concepts, students would do better overall even if we didn't teach every curriculum lesson. She was right—our state test scores went up even though we 'fell behind' in the curriculum."

The Winooski case provides particularly strong evidence for the importance of administrative support because it holds constant many variables (same district, same professional development, same coaching support, same student demographics) while varying the degree of administrative backing for prioritizing depth over coverage. The correlation between administrative support and student outcomes across grade levels suggests this factor substantially affects implementation success.

Evidence from Worcester County: Worcester County's sustained six-year improvement coincided with system-level changes that aligned accountability structures with the instructional model. District documents and administrator interviews revealed that in Year 2 of implementation, the district revised its benchmark assessment system to focus on High Leverage Concepts rather than textbook scope and sequence.

As the district mathematics supervisor explained in interviews: "We realized our benchmark tests were undermining the instructional model. Teachers were trying to implement Math Menu and focus on concepts, but then we were testing them on whether they'd taught page 47 of the textbook. So we rebuilt our benchmarks around place value, multiplicative reasoning, fractional reasoning—the High Leverage Concepts. Now teachers could slow down and go deep because the accountability system matched the instruction we were asking for."

District pacing guides were similarly revised in Year 3, with suggested time allocations increasing for High Leverage Concepts and decreasing for peripheral topics. For example, the third-grade pacing guide changed from allocating equal time across 180+ distinct standards to identifying 12 High Leverage Concepts with flexible timelines: "Develop multiplicative reasoning through equal groups, arrays, and area models: 4-8 weeks depending on formative assessment."

The systemic alignment of curriculum expectations, pacing guides, and assessment systems with the instructional model likely contributed to Worcester County's sustained improvement over six years. When accountability structures support rather than undermine the instructional approach, teachers face less tension between competing demands.

Evidence from Thompson School District: District-level leadership decisions in Thompson created conditions for model implementation. At the start of Year 1, district leadership made explicit decisions to prioritize conceptual depth, as documented in district communications to principals and teachers.

A district memo from the Assistant Superintendent for Curriculum and Instruction (distributed at the start of Year 1) stated: "This year, our focus is building strong mathematical thinking through High Leverage Concepts. Teachers are expected to spend sufficient time on foundational concepts to ensure students develop deep understanding. This may mean not completing every lesson in adopted curriculum materials. Principals should support teachers in making instructional decisions based on formative assessment rather than rigid adherence to pacing guides."

District mathematics coordinator interviews (n=3) indicated that this clear message from district leadership reduced building-level tension. As one coordinator explained: "Teachers need permission to slow down. When that permission comes from the superintendent's office, not just the math coach, it carries weight. Teachers knew they wouldn't be evaluated negatively for being 'off pace' if they were making data-based decisions about students' conceptual needs."

The consistency of improvement across all 23 Thompson schools—despite varying local contexts, demographics, and teaching staff—suggests that system-level support for the instructional model enabled broad implementation. When individual teachers or schools must navigate tensions between competing demands, implementation becomes fragile and dependent on local conditions. When district systems align to support the model, implementation becomes more robust and less dependent on individual principal or teacher initiative.

Patterns Across Sites: The pattern across sites is consistent: districts with stronger qualitative and quantitative outcomes demonstrated administrative and system-level support for instructional model implementation. This support manifested through:

- Explicit permission for teachers to deviate from pacing guides based on formative assessment
- Alignment of benchmark assessments to High Leverage Concepts rather than textbook scope and sequence
- Revision of pacing guides to allow flexible timelines for conceptual development
- Principal participation in professional development to understand instructional model rationale
- District-level communications that prioritize conceptual depth over content coverage

Sites with weaker outcomes or variable implementation showed patterns of:

- Continued pressure to maintain curriculum pacing despite student need for additional time on concepts
- Benchmark assessments that evaluated textbook coverage rather than conceptual understanding
- Individual teachers feeling isolated in implementing practices that differed from building or district norms
- Administrators who did not participate in professional development and therefore could not evaluate whether teaching practices aligned with the model

The qualitative evidence from these cases suggests that instructional change requires alignment of multiple system components. Professional development alone appears insufficient when accountability structures, curriculum expectations, and administrative evaluation criteria pull in different directions. Sites achieving stronger outcomes typically addressed these systemic alignment issues, creating conditions where implementing the instructional model aligned with rather than conflicted with other professional expectations.

Patterns and Interpretation Over Multiple PD Sites

Pattern 1: Sequence of Change Timeline

The most consistent pattern across sites is a sequence where teacher-reported instructional changes occur in Year 1, with measurable student achievement gains emerging in Year 2. This pattern held across different assessment types and demographic contexts.

Evidence from Thompson School District: The iReady data demonstrates this pattern clearly. Between the baseline year (2022-23) and Year 1 of implementation (2023-24), elementary schools showed mean score increases ranging from 16.3 points (Lincoln Elementary) to 27.9 points (Cottonwood Plains Elementary). However, between Year 1 and Year 2 (2024-25), gains continued at most schools. For example:

Table 01: Gains in TSD over Two Years

	Year 1 (Gain in points)	Year 2 (Gain in points)
Berthoud Elementary	23.5	22
Carrie Martin Elementary	21.7	22.1
Ivy Stockwell Elementary	19.9	24.9
Truscott Elementary	36.4	21.6

The pattern of sustained gains in Year 2 occurred at 18 of 23 schools, suggesting that initial instructional changes in Year 1 continued to translate into measurable achievement outcomes. Middle schools showed similar patterns but with more modest effect sizes (partial eta squared 0.19-0.70 versus 0.66-0.91 for elementary schools), suggesting that either the intervention is more effective at elementary levels or that middle school implementation faced greater challenges. The improvements in middle school achievement were statistically significant and educationally valuable, even if the effects were not as great.

Evidence from Winooski: The Grade 6 cohort data explicitly demonstrates this pattern of progressive achievement. Between 2023 and 2024 (first implementation year), this cohort's mean scale scores increased from 1,622 to 1,638—a modest 16-point gain. However, between 2024 and 2025 (second year), scores increased from

1,638 to 1,656—an 18-point gain with statistical significance emerging more strongly in the later period. The Grade 7 cohort, which had two full years of ALN-trained instruction by 2024, showed the strongest overall growth ($p < .001$, effect size 0.619).

Qualitative Evidence: Teacher survey data from multiple sites documented the pattern that achievement gains are seen after a minimum two years of implementation. In Plymouth Public Schools, first-year teacher surveys indicated that 73% of teachers reported changing their differentiation practices, but classroom observation data showed variable implementation quality. By Year 2, both survey data and observation protocols indicated more consistent, sophisticated implementation of Math Menu structures and formative assessment protocols.

Similarly, in Keene School District, the district's strategic decision to invest Year 1 in infrastructure building rather than immediate classroom implementation may have compressed the typical timeline, resulting in the unusual finding of 24% district-wide proficiency gains in what was effectively their Year 1 of full classroom implementation.

Cross-Site Consistency: This two-year pattern appeared regardless of assessment instrument (iReady diagnostic, state PARCC assessments, VTCAP, New Hampshire SAS) and regardless of demographic context (rural Vermont supervisory unions, suburban Massachusetts districts, large Colorado district). The consistency across diverse contexts strengthens confidence that the pattern represents a genuine implementation trajectory rather than artifact of specific measurement or population characteristics.

Thompson School District data showing year-over-year gains across three consecutive years suggests continued improvement beyond initial implementation. However, determining whether Year 3 gains represent continued fidelity improvement, teacher expertise development, or cohort effects (students who have experienced ALN-trained instruction for multiple years) requires additional analysis that disaggregates by student exposure duration.

Pattern 2: Disproportionate Gains for Historically Underserved Students

Perhaps the most notable finding is consistency of larger gains among student populations that typically show smaller responses to instructional interventions. Across multiple sites using different assessments, students receiving special education services, students from economically disadvantaged backgrounds, and English language learners demonstrated accelerated growth.

Evidence from C.P. Smith Elementary: Students receiving special education services at this Vermont elementary school achieved VTCAP proficiency rates approximately three times the state average for comparable populations. Specifically, while Vermont's statewide proficiency rate for students with IEPs in mathematics was approximately 18% during the evaluation period, C.P. Smith's students with IEPs achieved proficiency rates exceeding 50%. The school serves a high-poverty population (62% economically disadvantaged), yet outperformed wealthier Vermont schools on state assessments following ALN implementation.

Evidence from Keene School District: The disaggregated data reveals striking differential impact:

- District-wide proficiency increase: 24%
- Students receiving special education services: 16% increase
- Students from economically disadvantaged backgrounds: 35% increase

The 35% proficiency increase for economically disadvantaged students—substantially exceeding the already-strong 24% district-wide gain—demonstrates that the intervention disproportionately benefited students who typically show the smallest responses to instructional reforms. This pattern contradicts the common finding that educational interventions show largest effects for already-advantaged populations.

Evidence from Worcester County: The longitudinal data spanning 2017-2023 provides particularly compelling evidence of sustained impact for vulnerable populations. Six years after initial intervention, Worcester County Public Schools maintained mathematics performance exceeding state averages by 23 percentage points overall.

More significantly, the district's performance gaps between high-poverty schools and low-poverty schools narrowed substantially. Pocomoke Elementary, identified as the district's highest-poverty school, increased proficiency by 30 percentage points in the first implementation year and maintained performance at or above district averages in subsequent years.

The district's sustained outperformance for both students in poverty and students with learning differences (relative to state benchmarks) persisted despite complete turnover in some teaching staff and changes in district leadership, suggesting that the improved outcomes for vulnerable populations became institutionalized rather than dependent on specific individuals. This seems due to persistence of best practices introduced by ALN. Participants report that they continued to use practices like Math Menu because they “engaged the students” and “just seemed to work.”

Evidence from Winooski Public Schools: Winooski provides critical evidence because of its demographic composition. Winooski School District is home to 44% English language learners and 68% economically disadvantaged students. The district serves Vermont's most linguistically and economically diverse population. The Grade 7 cohort, showing the strongest growth (effect size 0.619, $p < .001$), was comprised predominantly of English language learners and students from refugee/immigrant families.

Disaggregated analysis revealed that English language learners in the responding cohorts (Grades 6 and 7) demonstrated statistically significantly greater growth than native English speakers. This finding is particularly notable because mathematics instruction often assumes linguistic facility, and ELL students typically show smaller gains on mathematics assessments due to language demands of word problems and mathematical discourse. The fact that Winooski's ELL population showed improved, rather than diminished, growth suggests that ALN's emphasis on multiple representations (concrete manipulatives, visual representations, abstract symbols) may reduce language barriers to mathematical understanding.

Evidence from Thompson School District: While the published analysis does not disaggregate by student subgroup, the range of effect sizes across schools provides suggestive evidence. Schools serving the highest-poverty populations in the district (Lincoln Elementary, Truscott Elementary, Laurene Edmondson Elementary) showed effect sizes at the high end of the range:

- Truscott Elementary: partial eta squared 0.79-0.85 (highest-poverty elementary)
- Lincoln Elementary: partial eta squared 0.67-0.79
- Carrie Martin Elementary: partial eta squared 0.83-0.88

These effect sizes equal or exceed those at more affluent schools, suggesting that the intervention was at least as effective—and possibly more effective—in high-poverty contexts.

Gains for historically marginalized students merits particular attention as it runs counter to common patterns where instructional reforms show largest gains for already-advantaged students. The mechanism likely relates to ALN's emphasis on making use of student thinking and understanding, rather than relying on a single approach, and differentiated access points for a diverse group of learners.

The emphasis on conceptual models, and representations before abstract symbolic manipulation may reduce barriers for English language learners who can demonstrate mathematical reasoning through manipulatives and drawings before

having complete facility with English mathematical vocabulary. Teacher interview data from multiple sites noted that students who previously disengaged during whole-group instruction participated more actively during Math Menu sessions where they received instruction matched to their conceptual level.

The focus on High Leverage Concepts rather than broad coverage may also benefit students with learning disabilities. Rather than falling progressively further behind as grade-level curriculum races ahead, students working on foundational multiplicative reasoning or additive reasoning concepts, even when chronologically in upper elementary grades, build conceptual foundations that enable later access to grade-level content. Multiple special education teachers across sites reported that this approach reduced their reliance on purely procedural compensatory strategies, instead building genuine mathematical understanding.

Pattern 3: Implementation Differences Affected Outcomes

The Winooski case, showing differential outcomes across grade cohorts within a single district, illustrates how implementation conditions affect results. Even within a single district with consistent professional development, outcomes varied substantially based on grade-level implementation factors.

Evidence from Winooski's Grade-Level Variation: The within-district variation provides natural experiment conditions:

Grade 7 cohort (started ALN instruction in Grade 5):

- Mean scale scores: 1,654 (2023) → 1,686 (2024) → 1,714 (2025)
- Effect size: 0.619 (large)
- Statistical significance: $p < .001$
- Two full years of ALN-trained instruction by final measurement

Grade 6 cohort (started ALN instruction in Grade 4):

- Mean scale scores: 1,622 (2023) → 1,638 (2024) → 1,656 (2025)
- Effect size: 0.545 (large)
- Statistical significance: $p < .001$
- Largest gains between Year 2 and Year 3

Grade 4 cohort:

- Mean scale scores: No statistically significant change
- Implementation challenges documented in teacher surveys
- One year of ALN-trained instruction

Qualitative data from teacher surveys and administrator interviews revealed that Grade 4 teachers faced curriculum alignment challenges, with district-adopted mathematics curriculum materials conflicting with ALN's High Leverage Concepts approach. Teachers in the elementary school (preK-5) work under a different principal, in a different building, and with a different culture than teachers in the middle school (Grades 6-8). Additionally, Grade 4 had higher teacher turnover during the implementation period. In contrast, Grades 5-7 teachers reported strong administrative support for prioritizing conceptual depth over curriculum pacing guides, and these grades maintained stable teaching teams.

Evidence of Coaching Impact: Sites with stronger outcomes consistently featured sustained embedded coaching beyond initial workshop professional development. Thompson School District's district-wide consistency (23 of 23 schools showing significant improvement) occurred in context of*:*

- Monthly embedded coaching visits to every school
- Building-level mathematics instructional coaches trained in ALN methods
- District mathematics coordinators observing classrooms and providing feedback
- Regular collaborative teacher learning sessions focused on student work analysis

In contrast, sites with more variable outcomes typically provided workshop-based professional development without systematic coaching infrastructure. The differential outcomes between Winooski's responding grades (which had consistent coach presence) and non-responding grades (where coaching was less intensive due to scheduling constraints) supports this pattern.

Evidence from Keene's Infrastructure Investment: Keene School District's strategic decision to invest the first full year (2022-23) in infrastructure building provides quasi-experimental evidence about coaching importance. Rather than immediate classroom implementation, the district:

- Developed a cohort of 12 instructional coaches (roughly 1 coach per 250 students, and 14 teachers)
- Conducted comprehensive needs assessment
- Aligned assessment systems and curriculum materials
- Built administrator understanding of instructional model

This year-long investment preceded the district's first year of broad classroom implementation (2023-24), during which they achieved the atypical result of 24% district-wide proficiency gains in what was effectively their Year 1. The finding

suggests that robust infrastructure—particularly coaching capacity—may compress the typical two-year timeline for achievement gains, though determining causality requires comparison conditions.

Evidence of Administrative Support: Sites with stronger outcomes featured administrative backing for instructional changes that sometimes contradicted existing policies or materials. In Worcester County, building principals attended ALN professional development alongside teachers and participated in classroom learning walks focused on identifying ALN practices. Principals reported understanding the model sufficiently to discuss High Leverage Concepts, Math Menu structures, and formative assessment protocols in teacher evaluation conferences.

Teacher interview data from multiple sites revealed tension between ALN's depth-over-coverage approach and district pacing guides derived from textbook sequences. Teachers who successfully implemented the model reported administrator support for deviating from pacing guides when students needed additional time on foundational concepts. For example, one third-grade teacher in Orange East Supervisory Union described spending six weeks on multiplicative reasoning (versus the curriculum's two-week allocation), with principal support, resulting in students demonstrating stronger performance on later fraction and multi-digit multiplication units that rely on multiplicative understanding.

In contrast, Grade 4 teachers in Winooski (the non-responding cohort) reported feeling pressured to maintain curriculum pacing despite student need for extended time on place value concepts, citing district benchmark assessments aligned to textbook scope and sequence rather than High Leverage Concepts framework.

Evidence of Teacher Understanding: Qualitative data suggests that outcomes correlate with teacher understanding of conceptual rationale rather than procedural compliance. In Plymouth Public Schools, researchers coded teacher interview transcripts for depth of understanding about why ALN practices might improve student learning. Teachers coded as "high understanding" (those who articulated connections between practices and learning theory) showed stronger implementation fidelity on classroom observation protocols than teachers coded as "procedural compliance" (those who could describe what to do but not why).

For example, "high understanding" teachers described Math Menu as enabling them to provide just-right instruction based on formative assessment, with groupings designed to address specific conceptual gaps revealed through work sorts. "Procedural compliance" teachers described Math Menu as ability grouping with rotation through stations. Observation data indicated that the former group used formative assessment to drive grouping decisions and adjusted instruction based on

student responses, while the latter maintained relatively static groups and delivered similar instruction regardless of student response patterns.

Evidence of Sustainability: Worcester County's sustained six-year improvement (2017-2023) and Thompson's district-wide consistency suggest that when districts build internal capacity through trained coaches and establish new instructional norms, improvements may persist beyond active external support periods.

Worcester County's longitudinal trajectory shows:

- Year 1 (2017-18): Large gains during intensive ALN coaching
- Years 2-3 (2018-20): Maintained gains with reduced external support
- Years 4-6 (2020-23): Continued to exceed state averages by 23 percentage points with minimal external support

To continue success, the district maintained:

- Cohort of trained building-level coaches who had participated in ALN professional development
- Administrator expectation that mathematics instruction would include Math Menu and formative assessment protocols
- Common language around High Leverage Concepts in grade-level planning in grades K-5
- Modified district benchmark assessments to align with conceptual framework

The sustained outcomes six years post-intervention, despite staff turnover and leadership changes, suggests successful transition from external support to internal capacity. However, the absence of detailed implementation fidelity data from 2020-2023 prevents determining which practices persisted with high fidelity versus which may have faded while still producing sustained outcomes.

Thompson School District's consistency across 23 diverse schools in Year 1-3 may indicate early sustainability factors, though the district continues to receive external support. The uniformity of outcomes across schools serving vastly different populations (affluent suburban to high-poverty contexts, small rural schools to large comprehensive schools) suggests establishment of district-wide norms and expectations rather than individual school initiative, which theoretically supports longer-term sustainability.

Limitations and Methodological Considerations

This synthesis has significant limitations that constrain causal inference:

Absence of Comparison Groups: No studies employed randomized control designs or matched comparison schools. When improvement occurs, multiple concurrent changes (curriculum adoption, leadership changes, demographic shifts, state policy changes) could contribute to outcomes. The consistency of improvement across diverse contexts strengthens confidence, but does not substitute for experimental design.

Selection Effects: These cases represent districts that chose to partner with ALN, often because they identified significant needs. Districts with different characteristics, resources, or readiness levels might experience different outcomes. We cannot determine how results compare to similar districts selecting different professional development approaches.

Variable Data Quality: Some sites provide rich qualitative data with limited quantitative outcomes. Others offer test score data without detailed implementation information. Few include classroom observation data that would illuminate specific practices driving results. This variation limits the ability to specify necessary versus sufficient model components.

Implementation Fidelity Measurement: Most sites lack systematic fidelity measurement. We observe that teacher practice changed and student achievement improved, but cannot fully specify dosage-response relationships or which model components are essential.

Publication Bias: These reports document partnerships where data was collected and analyzed. We lack information about sites where ALN worked but outcomes were not evaluated or where partnerships concluded without formal impact analysis. This may lead to overestimation of typical effects.

These data suggest ALN's approach produces meaningful improvement across diverse contexts, with effect magnitudes exceeding what would be expected from measurement error or regression to the mean. The consistency of findings across different assessment instruments, demographic contexts, and geographic regions strengthens confidence. However, without more rigorous designs including comparison conditions and systematic fidelity measurement, definitive causal claims remain premature.

Implications for Practice and Policy

Several implications emerge for district leaders considering mathematics instructional improvement initiatives:

Timeline Expectations: Districts should anticipate instructional practice changes within one year, with measurable student achievement gains typically emerging in Year 2 of sustained implementation. The Keene case suggests infrastructure investment before broad implementation may accelerate this timeline, though such approaches require substantial upfront resource commitment.

Internal Capacity Development: Worcester County's six-year sustained improvement and Thompson's district-wide consistency point toward the importance of internal capacity building. Districts should plan from implementation initiation for how improvement will be maintained beyond external support periods. This includes developing internal coaching expertise, aligning assessment and curriculum systems with instructional models, and ensuring administrators understand the model sufficiently to support continuation.

Focus on Achievement Gap Reduction: For districts where closing achievement gaps for students with disabilities, students in poverty, or English language learners represents a priority concern, the evidence pattern is particularly relevant. The consistent finding of disproportionate gains for these populations across multiple contexts and assessment types suggests the approach addresses systematic instructional inequities in traditional mathematics teaching.

Implementation Variability: Not every teacher will adopt practices at identical rates. The Winooski data demonstrates variable outcomes across grade cohorts within single districts. Leaders should plan for differential adoption rates and address implementation barriers rather than expecting uniform transformation. Strong outcomes appear associated with sustained coaching support, administrative backing for instructional changes, and teacher understanding of conceptual rationale rather than procedural compliance.

Systematic Evaluation Planning: Several promising cases in this synthesis lack robust quantitative outcome data, limiting conclusions about effectiveness. Districts should plan from implementation start for measuring both implementation fidelity (are teachers changing practice?) and student outcomes (are students learning more?). This requires baseline data collection, systematic ongoing measurement, and ideally, comparison conditions that allow stronger causal inference.

Conclusion

This evaluation synthesizes evidence from ten partnerships between All Learners Network and school districts or supervisory unions. The most robust quantitative evidence comes from Thompson School District, where all 23 schools showed statistically significant improvement with generally large effect sizes across three implementation years. Supporting evidence from smaller studies in Vermont, New Hampshire, Maryland, and Massachusetts shows consistent patterns: instructional change preceding achievement gains, disproportionate improvement for historically underserved students, and sustainability when districts build internal capacity.

The evidence base has clear limitations. Absence of randomized designs or matched comparison groups limits causal claims. Variable data quality across sites constrains understanding of implementation mechanisms. Publication bias may result in overestimation of typical effects.

Despite these limitations, several findings merit attention. The magnitude and consistency of achievement gains, particularly for students with disabilities and students from economically disadvantaged backgrounds, exceeds what measurement error or regression to the mean would predict. The Thompson School District evidence, with consistent improvement across 23 diverse schools, provides particularly strong suggestive evidence of positive intervention effects. The pattern of instructional change preceding achievement gains aligns with theoretical predictions and has face validity.

For districts facing persistent mathematics achievement gaps, particularly for students receiving special education services or students from economically disadvantaged backgrounds, the accumulated evidence suggests ALN's approach warrants serious consideration. However, districts should maintain realistic expectations about implementation timelines, commit to building internal capacity for sustainability, and develop their own evaluation plans for measuring progress.

The central hypothesis, that all students can develop mathematical thinking when instruction is grounded in conceptual understanding, differentiated based on formative assessment, and focused on High Leverage Concepts, receives support from this evidence synthesis. However, the hypothesis requires testing through more rigorous evaluation designs including comparison conditions and systematic implementation fidelity measurement.

All Learners Network has engaged program evaluators for comprehensive analysis across current partner districts during 2024-2025. This work will provide more systematic evidence about implementation conditions, successful practices, and overall impact magnitude.

Data Sources

This evaluation synthesizes findings from impact reports and analyses conducted in:

Thompson School District, Colorado (2022-2025) | Worcester County Public Schools, Maryland (2017-2023) | Winooski Public Schools, Vermont (2023-2025) | Keene School District, New Hampshire (2022-2024) | Orange East Supervisory Union, Vermont | Plymouth Public Schools, Massachusetts | C.P. Smith Elementary School, Vermont | Union Street School, Vermont | Central Vermont Supervisory Union | Franklin West Supervisory Union, Vermont

Complete impact reports and technical documentation available upon request from All Learners Network.