



Directors, Not Users

A Framework for AI-Integrated Instruction
and Professional Development in Massachusetts

Mike Luekens

October 2026

Directors, Not Users

A Framework for AI-Integrated Instruction and Professional Development in Massachusetts

Mike Luekens | DirectEd | October 2026

Foreword

I texted my sons recently, “What are three important, realistic things a teacher could do differently to help you or kids your age learn better?” They’re 19 and 16, and grew up in MA K-12 public schools, as I did. My oldest quickly fired back three ideas:

1. “Having two teachers in the room is a huge help to everyone’s learning”
2. “Teachers providing sheets that are designed for taking notes for the lesson that day”
3. “Providing a schedule to everyone of the due dates of projects/assignments/exams at the start of the year”

Let’s start with #2 and #3, since they’re easier. A good, thoughtful syllabus, refined every semester, with materials for note taking pre-developed and distributed to students each week, seem like pretty reasonable requests to me. But doing this type of planning work well takes both time and skill. On top of everything else, planning and prep work takes up *a lot of time*. Time that teachers never have enough of.

Now onto his #1 item, which is funny because I asked about things “a” teacher could do...

Would two teachers in every classroom help with planning and prep? Absolutely! Could it help improve learning for the students in those classrooms? Of course! Is it feasible to think we’d ever get to a point where all communities can afford to put two teachers in every classroom? Of course not! Not even close. Different types of team or shared teaching can absolutely contribute to better learning. But any practical education administrator knows that putting two teachers in classrooms is prohibitively expensive at any type of scale, with a few exceptions.

It would be easy to dismiss the suggestion – it’s just not realistic. But it begs other important questions:

- “How can we provide **more support** to every single teacher in every single classroom, in a way that’s far more efficient and cost-effective than hiring?”
- “How can we **empower** every single teacher to focus **more time on instruction**, using the same tools and resources that are already available to us?”

My younger son followed up with his ideas a few hours later:

1. “Building better connections with the students; when students have good connections with teachers they care to listen and learn more”
2. “Behavior should be part of the class grade because if you’re focused on not getting distracted or talking to people you will be more focused on learning”
3. “More class participation so everyone can be involved more”

Again, these suggestions make sense, but all require time, effort, and skill by teachers. They also raise similar questions: How can we provide more support to teachers, allowing them to focus more energy on building connections with students, addressing behavioral obstacles, and developing students to actively participate and take on a greater level of involvement and ownership in their own learning?

“Can AI help?”

“Sure, AI can probably help.”

“Ok, **but how?**”

This document lays the research foundation and a framework to answer that question: “**but how?**”

AI, when used well, can be a force multiplier in education. It can leverage teachers’ capabilities and support more and better time focused on what actually matters in the classroom: teaching, learning, and student outcomes.

Executive Summary

Educators and researchers have known for a long time what improves teaching. But scaling effective teaching practices across entire districts is, to put it mildly, challenging. Instructional coaching has some of the strongest causal evidence in education research: a meta-analysis of 60 causal studies found effects of 0.49 standard deviations on teachers' practice and 0.18 on student achievement (Kraft, Blazar, and Hogan, 2018). But good coaches are hard to find and expensive, and the effects get smaller as programs get bigger. Meanwhile the one-day workshop, still the most common form of PD, scales easily and rarely produces lasting change in the classroom. So for years the approaches that work have been hard to scale, and the approaches that scale haven't worked.

This challenge has only become more pressing. Of roughly 400 Massachusetts districts, just 14 are back to pre-pandemic MCAS performance in both ELA and math, and the federal recovery money is spent. Teachers are picking up AI on their own, even though 68 percent of teachers haven't participated in any type of AI training with their school or district. The free educator versions of the big AI platforms end in June 2027. If adopted, the Graduation Council's recommended requirements, including an AI and digital literacy milestone, would begin with ninth graders entering in fall 2027. Districts are going to adopt AI one way or another. What's still open is whether it improves teaching or just speeds up whatever is already happening.

The core idea. Take the practices teachers already know—structured discussion, scaffolding, formative assessment, specific feedback. For those, the thing holding teachers back is time. Each one takes hours of preparation most teachers simply don't have, so they show up in some lessons and not others. AI that's been set up in advance with a district's adopted curriculum, state standards and the district's own improvement priorities can take on much of that preparation. It doesn't give teachers expertise they lack. It frees up the expertise they have. For practices teachers know less well, the constraint is knowledge and skill, and that's what coaching is for.

A multiplier multiplies weak choices too, and AI can wear down judgment as easily as it builds it. So the framework doesn't start with the tool. It starts with evidence about where a district's students are falling behind. It builds on the curriculum the district already adopted instead of working around it, and it develops teachers' judgment about what their students need next. A user takes what the tool hands back. A director decides what the class needs and makes sure that's what gets built. Helping teachers become directors in leveraging AI tools is the whole point of the professional development approach.

How it works. The framework runs on a continuous improvement cycle (Diagnose-Respond-Measure-Adjust) drawn from improvement science:

- **Diagnose.** A peer-benchmarked diagnostic compares the district with similar districts by grade, subject, school and student group, using multi-year data to find gaps and strengths that state reports don't show.
- **Respond.** A strategy prescription matches those gaps to three to five instructional practices with strong research behind them.
- **Equip.** Each participating teacher gets an AI planning assistant configured ahead of time with the district's adopted curriculum, the standards and the prescription. Because the prescribed practices are built into the assistant's training, the professional development doesn't stop when the PD day ends.
- **Measure and adjust.** A Build Day, monthly team coaching cycles led by the district's own coaches, and a monthly community of practice build teachers' judgment over the year. Quarterly reviews look at classroom practice, time use, the quality of what the assistant produces and student results against the original diagnostic, and the prescription gets adjusted.

Equity is part of the structure. Every measure is broken out by race, income, English learner status and disability, and every review asks "for whom?" before it asks "is it working?" The assistant defaults to giving students access to grade-level work rather than handing them easier work. That matters, because schools serving more disadvantaged students are integrating AI more slowly (Campos and Singleton, 2026). Left alone, the achievement gap turns into an AI gap.

What the evidence says. The instructional practices, curriculum-focused professional development and the improvement method all have strong support. The early randomized research on AI that supports adults (rather than replacing them) is encouraging. Tutors with AI support produced more student mastery, especially the least experienced tutors (Wang et al., 2024), and teachers using AI to plan lessons cut planning time by about 31 percent with no loss in quality (Education Endowment Foundation, 2024). Some evidence points the other way, too. Open-ended AI access can raise immediate performance while weakening learning (Bastani et al., 2025), and people who lean on capable automation tend to miss its mistakes. Whether the full system raises achievement hasn't been shown yet. The numbers to beat are about 0.14 standard deviations for a year of student-facing AI tutoring and about 0.18 for coaching.

How we'll know. I'll register the evaluation design publicly before the first engagement starts. It uses staggered starts and comparisons with peer-district trends, corrects for regression to the mean, and measures time use and output quality directly instead of relying on self-report. No single district is big enough to detect effects of this size, so each district's results are reported descriptively. The real test pools participating districts as engagements accumulate, analyzed against the registered plan. Costs are reported alongside effects, and disappointing results get published along with the good ones.

What districts can count on. The framework doesn't cut teaching or paraprofessional positions. Observation data is never used in educator evaluation, and that's agreed with the local association before anyone observes a class. Student records stay in district systems, and only accounts covered by a signed Massachusetts Student Data Privacy Agreement are used. The district owns everything built for it. I take no payments from AI companies, publishers or platform providers.

Who does the work. The district does, and my job is to build its capacity to do it well. I build the diagnostic, the prescription and the first version of the assistant, design and lead Build Day, and train the district's instructional coaches and working group to run the coaching cycles, the community sessions and the quarterly reviews. They run that work in their own schools from the start, with my support, and I step back over the course of the year. By year end, the district is running the cycle on its own.

This paper lays out the framework, the research behind it, and details on what an engagement looks like.

Introduction

One of the books that stuck with me from graduate school is Frederick Hess's *Spinning Wheels: The Politics of Urban School Reform* (1998). Hess argued that district leaders cycle through largely symbolic initiatives to show they're "making a difference," while the slower, more focused work of building expertise in the teaching practices that are actually tied to results gets neglected.

Twenty-eight years later, the pattern Hess described has played out on a national scale. No Child Left Behind, Race to the Top, Common Core – each one changed something structural: standards, tests, accountability, governance. Each asked something real of the people who carried it out. And each left the fundamental challenge largely untouched, which is what happens between a teacher and students in a classroom on a weekday morning.

The results are hard to argue with. On the 2024 Nation's Report Card, 12th-grade reading fell to its lowest level since the assessment began in 1992, and 12th-grade math to its lowest since 2005. Compared with 2022, no state saw reading gains in either fourth or eighth grade (National Center for Education Statistics, 2025). U.S. math performance on the 2022 PISA was among the lowest the country has recorded (OECD, 2023). Meanwhile, inflation-adjusted spending reached \$20,322 per student in 2023, about 36 percent more than in 2002, and close to \$1 trillion in total (Reason Foundation, 2025).

What strikes me most is what has happened to the word "development." Professional development has turned into a compliance activity – points, hours, certificates – rather than a process that actually develops professional capability. On the student side, achievement gets measured by a single score on a single test. Few systems track whether students are developing the capacity to think, reason, evaluate and direct their own learning, which are the capabilities that will matter most in a world where AI handles information retrieval. Development, in both senses, ought to mean what it says.

I've been inside this system. At AIR, I managed the Schools and Staffing Survey, part of the data infrastructure behind federal education policy. At Measured Progress, I led work on a statewide high-stakes assessment program of the kind used to judge whether reforms worked. On my city's School Committee, I chaired the subcommittee responsible for curriculum, instruction and accountability. As a parent I've seen firsthand how the pandemic impacted learning for my own kids.

I also spent 17 years outside education, running large operations as an executive at Liberty Mutual, and that work taught me many things that apply directly here. A judgment is only as good as the comparison behind it: nobody in insurance prices a risk against the average of every risk, yet education compares every district to the state average. Giving people better tools fails

unless you develop the skill and judgment to use them, which is exactly what's happening now with AI in schools. And we don't know whether something is working until we measure it. Closing that measurement loop is how organizations actually improve.

AI is the most likely candidate for the next spinning wheel. It's new and highly visible, it demos well, and nearly every education vendor now sells some version of it. Districts will be under pressure to adopt something quickly, announce it and move on. This framework is built to avoid that pattern. It's not intended as yet another "reform." Instead it works inside the existing structure – existing curricula, existing teachers, existing schedules. It treats AI as a way to make known, effective teaching practices easier to carry out every day, and then it measures whether instruction actually improves and whether students actually develop. That measurement isn't a compliance exercise. It's the feedback loop that lets the system correct itself.

Part I: Where Districts Stand

Massachusetts Results

DESE released the latest MCAS results on September 22, 2026. Forty percent of students in grades 3–8 met or exceeded expectations in English Language Arts – down two points from 2025 and twelve from 2019. Math was 41 percent, flat from the year before and eight points under pre-pandemic levels. Grade 10 ELA fell six points in a single year and now sits sixteen points below 2019. Only 14 of approximately 400 districts are back to pre-pandemic performance in both subjects. And the \$2.6 billion in federal ESSER recovery funds is gone.

There are bright spots, and they're real. Fifty-six Schools of Recognition beat the statewide trend. Grade 8 math proficiency went up four points. Chronic absenteeism dropped to 17.1 percent from a peak of 22.2.

Every district has excellent teachers. The hard part is making excellent teaching the norm.

Professional Development Isn't Working

Massachusetts educators must earn 150 Professional Development Points every five years, and districts spend real state and federal money (Title II-A included) to support that. The problem isn't so much what gets spent as what it buys.

TNTP's study *The Mirage* (2015) looked at three large districts and a charter network and estimated they were spending roughly \$18,000 per teacher per year on teacher development, once staff time was counted. Only about three in ten teachers improved substantially, and no type of activity reliably predicted who would. The study has its critics. It measured improvement mostly through evaluation ratings and value-added estimates, and the cost figure depends on how you count staff time. Still, its central finding – spending didn't reliably buy improvement – lines up with the wider literature. In March 2026 the U.S. Government Accountability Office concluded that research on PD effectiveness is still mixed, and found that only 40 percent of teachers consider most of their PD a good use of their time (GAO-26-107874). The same report found that teachers rate collaborative learning with colleagues as the most useful kind. Darling-Hammond, Hyler and Gardner's (2017) review of rigorous studies points the same way: effective professional development is content-focused, uses active learning, supports collaboration, models effective practice, provides coaching and expert support, builds in feedback and reflection, and is sustained over time. DESE's own definition of high-quality professional learning says it should be sustained, job-embedded and connected to practice. The one-day workshop, which is still the most common model, is none of those.

Coaching does better. Kraft, Blazar and Hogan's (2018) analysis of 60 causal studies found coaching improved teachers' instructional practice by 0.49 standard deviations and student

achievement by 0.18, which is a large effect by education research standards. Two other findings matter for this framework. Effects were noticeably smaller in programs serving 100 or more teachers. And the review found no clear relationship between how many hours of coaching teachers got and how big the effect was, which suggests the focus and quality of coaching count for more than the volume. Good coaches are expensive, and there aren't many of them.

Teachers Aren't Waiting on AI

DESE has moved deliberately on AI. It convened an AI Task Force in August 2024, which issued recommendations that December. The department built resources through spring and summer 2025 – guidance for district leaders, quick reference guides, a planning and implementation companion, an AI Literacy for Educators course, a webinar series and facilitation guides – and is providing implementation support in 2025-26, with policy considerations to follow in 2026-27 (Massachusetts Department of Elementary and Secondary Education, 2025). What DESE has built is thoughtful and useful. What it doesn't provide, and isn't designed to provide, is implementation inside a district: a diagnostic, a strategy tied to specific gaps, tools configured for the district's curriculum, coaching and measurement. DESE tells districts what to think about. The question is who helps them do it.

The state also funds *Future Ready: AI in the Classroom*, a \$135,000 pilot run with Project Lead The Way. It offers a 50-hour professional learning experience for high school teachers in about 30 districts and focuses on teaching AI as course content (Massachusetts Executive Office of Education, 2025).

Teachers have gone ahead anyway. MagicSchool reports more than two million educator users. Claude for Teachers and ChatGPT for Teachers are free to U.S. educators through June 2027, and Google is building Gemini into Classroom and Workspace. Next door in New Hampshire, a statewide contract put Khanmigo in 50 districts.

Training hasn't kept pace. In the Gallup-Walton Family Foundation survey (2025), 68 percent of teachers said they hadn't taken part in any AI training provided by their school or district, and only 19 percent said their school had an AI policy. RAND's American School District Panel (2025) found the same gap looking at it from the district side.

New Graduation Requirements

In November 2024, Massachusetts voters got rid of MCAS as a graduation requirement. In June 2026, the K-12 Statewide Graduation Council delivered its final recommendations for what replaces it: MassCore coursework, a common assessment, a culminating experience such as a capstone or portfolio, and My Career and Academic Plan (MyCAP) milestones in work-based learning, financial literacy, and AI and digital literacy. The recommendations apply only if

adopted, and implementation timelines are still being set, but the state is strongly encouraging districts to meet MassCore for ninth graders entering in fall 2027. About 74 percent of districts already have local requirements that meet MassCore (K-12 Statewide Graduation Council, 2026). The AI and digital literacy milestone is the new element, and few districts are ready for it. Capstones ask for sustained, applied, evaluative work from students. And it's hard to see how teachers teach AI literacy well without building their own first.

The Curriculum Context

Massachusetts has spent the better part of a decade nudging districts toward high-quality instructional materials. DESE's CURATE reviews rate core programs against the state frameworks, and state grant programs have helped districts adopt the highly rated ones. The reasoning holds up. A strong curriculum puts coherent, grade-level materials in front of every student, no matter which teacher they happen to draw.

So any AI tool that makes it easy for teachers to churn out their own materials has to be judged against that investment. Part II gets into how this framework is built to support adopted curricula rather than compete with them.

What's Missing

Guidance isn't in short supply. About 40 states had issued official K-12 AI guidance or policy frameworks by September 2026 (Playlab, 2026), and there are solid frameworks from UNESCO, TeachAI and ISTE, Child Trends and DESE itself. They do a good job describing what good AI integration looks like. They say much less about how a district actually gets there.

Training isn't in short supply either. In 2025 the American Federation of Teachers launched a National Academy for AI Instruction with \$23 million in support from Microsoft, OpenAI and Anthropic (AFT, 2025). The platform companies run their own courses, and the state pilot reaches teachers in about 30 districts. What these efforts mostly don't do is tie AI to a specific district's curriculum and achievement data, or check whether it improves results there. National training teaches tools and general practices. The state pilot treats AI as a subject. And as of October 2026, very few of the roughly 60 providers on DESE's list of pre-qualified professional development vendors offer AI-integrated instructional support of this kind.

The gap is at the district level – AI connected to the curriculum teachers actually teach and the kids they actually have, with results measured well enough to justify next year's budget.

That gap is going to close one way or another. Districts that build their own capacity now will get to shape what AI-integrated teaching looks like in their schools. The ones that wait will probably end up with whatever shows up prepackaged.

Part II: Three Core Ideas

The Multiplier

Ask teachers what keeps them from teaching the way they know they should, and time comes up first. Teachers in RAND's 2024 State of the American Teacher survey reported working about 53 hours a week, roughly nine more than comparable working adults, and about 15 of those hours fell outside their contracted time (Doan et al., 2024).

Picture a teacher with five preps and 125 students. She knows scaffolded versions of a task would help her struggling readers. She knows an exit ticket sorted by misconception would sharpen tomorrow's lesson, and that a good structured discussion needs questions planned at several levels of difficulty. She also knows exactly how many hours each of those takes – and that there's a pile of parent emails, IEP progress notes and grading waiting behind them.

That preparation load is a bottleneck. It goes a long way toward explaining why practices with strong evidence show up in some lessons and not others, and in some classrooms and not others. None of these strategies are new. What's been missing is a practical way to use them day in and day out.

Time isn't the only constraint, though, and I don't want to pretend it is. What teachers know matters. Teachers' mathematical knowledge for teaching predicts their students' gains (Hill, Rowan, and Ball, 2005), and practices like formative assessment often get adopted more in name than in substance. If time were the only obstacle, all the hours districts already put into professional development would show more of a return than *The Mirage* found. So the framework treats the two constraints separately. Where teachers already know a practice well, the binding constraint is time, and the assistant goes after it. Where they know it less well, the constraint is knowledge and skill, and coaching handles that. Once coaching has built a new practice, the assistant helps make it sustainable.

AI that's been prepared ahead of time with a district's curriculum, standards and priorities changes the math on preparation. Differentiated materials take minutes instead of an evening. Exit tickets get sorted before the next class rather than at the end of the unit. Parent communications and first drafts of routine paperwork stop eating up the planning period. None of that gives a teacher expertise she doesn't already have; it takes away the things keeping her from using the expertise she does have. That's what I mean by a multiplier. Same skills, same judgment, same relationships with kids – used fully, in all of her classes.

Of course, a multiplier multiplies whatever you feed it. Apply AI to weak instructional choices and you get weak instruction faster. So the framework doesn't stop at the tool. The diagnostic decides where to aim, the prescription decides which practices to strengthen, and the professional development builds the judgment that decides what actually gets produced.

Built on the Adopted Curriculum

The quickest way for AI to undermine instruction would be to turn every teacher into a curriculum writer.

The research on teacher-sourced materials should give anyone pause. When the Thomas B. Fordham Institute had expert reviewers rate popular supplementary materials for high school English, including lessons from Teachers Pay Teachers, most were judged mediocre or not worth using (Polikoff and Dean, 2019). A tool that can generate unlimited supplementary material could reproduce that result faster, and pull teachers away from the curriculum their district paid for and trained them on.

The planning assistant is built to work the other way. Its curriculum layer is the district's adopted program – the unit sequence, the texts and tasks, the vocabulary, the common misconceptions. When a teacher asks for practice, the assistant starts from the lesson in the adopted curriculum and supports it with scaffolds, checks for understanding and discussion questions built around that lesson's tasks. It doesn't write replacement units. By default it sends teachers back into the curriculum, not around it.

The research here is actually encouraging. In a meta-analysis of 95 experimental and quasi-experimental studies of STEM instructional improvement programs, Lynch, Hill, Gonzalez and Pollard (2019) found an average effect of 0.21 standard deviations on student outcomes. Results were stronger for programs that trained teachers to use curriculum materials, focused on content and how students learn it, and gave teachers time to work through implementation problems together. Professional development organized around the curriculum teachers actually teach is one of the better-supported designs out there. The coaching and community sessions in this framework follow that design, and the assistant carries it into everyday planning.

The Agentic Harness

How much the multiplier delivers depends on what kind of AI teachers get. Most teachers using AI today are working with a general chatbot. It starts from scratch every session and knows nothing about the curriculum, the students' needs or the district's priorities. It does exactly what it's asked and nothing more. Getting good work out of it takes time and skill most teachers can't spare, which is one reason the time savings from AI have been so uneven so far.

People in AI engineering use the term *agentic harness* for something different: an AI system set up in advance with the knowledge a job requires, connected to the information it needs, and organized into structured workflows (instead of a blank chat window waiting for instructions). A harness already understands the work. A skilled professional can point it at a specific outcome, and it brings the context, standards and quality checks along.

The planning assistant at the center of this framework is an agentic harness for teaching – a coordinated set of prebuilt agents, each designed for one part of a teacher's work:

- **Lesson preparation:** warm-ups, lesson sequences and exit tickets aligned to the unit and standards.
- **Differentiation:** versions of a grade-level task at several levels of support, without lowering the bar.
- **Formative analysis:** exit tickets and quick checks sorted by standard and misconception, with suggestions for reteaching.
- **Discussion and questioning:** discussion protocols and question sequences that move from literal to analytical to evaluative.
- **Feedback:** rubric-aligned draft feedback on student work for the teacher to review and edit.
- **Family communication and paperwork:** parent messages in the teacher's voice, first drafts of IEP progress notes, conference preparation and logistics.

All of the agents draw on the same foundation: the Massachusetts standards, the district's adopted curriculum, the diagnostic findings and the strategy prescription, plus instructions to check for alignment, accuracy and grade-level rigor. So when a teacher asks for practice on tomorrow's lesson, the agent built for that job already knows the unit, the standards, where her students need help, and that scaffolds should lead up to grade-level work. The goal is for teachers not to need any prompting skill, and for even a vague request to come back aligned to the right unit. I don't want to just assume it does that. The quarterly output audits described in Part VI will measure how well it holds up, and Appendix C lays out how the harness is built.

The harness also carries the district's prescription into day-to-day work. Say the diagnostic calls for structured discussion in Grade 4 ELA. The discussion agent for Grade 4 teachers then defaults to discussion protocols built around the targeted standards, and each time a teacher uses it, the prescribed practice comes built into what she gets back. That's how the professional development keeps going after the PD day is over.

Put simply, the agents do the preparation, and the teacher directs them. The harness makes the multiplier possible; the teacher is the reason it's worth having.

Directors, Not Users

This is the idea the framework is named for, and it's easy to misread.

It has nothing to do with prompting. The agents are prebuilt precisely so teachers don't need prompting skill. With the curriculum, standards and district priorities already loaded, a teacher who types "practice for tomorrow's lesson" should get material aligned to the right unit and standards, at several levels of support, in a couple of minutes. The assistant is also designed to

fill in vague requests on its own, adding alignment and differentiation the teacher didn't ask for. The intelligence lives in the preparation, not in how the request is worded.

Most AI training teaches people to use a tool: where to click, what to type, what comes back. That gets you functional competence, which is a low ceiling. Someone who learns to use a word processor can produce documents. Someone who learns to write can change how readers think. Same tool, entirely different skill.

What separates a director from a user is professional judgment. A user asks for a quiz and uses whatever comes back. A director starts with what her students need – the standard they haven't mastered yet, the misconception last week's exit ticket turned up, the level of thinking the unit calls for. She picks the instructional move that fits, has the assistant produce the material, and then judges it the way she'd judge a colleague's work. Is it accurate? Is it aligned? Did the differentiation keep the rigor of the task, or quietly lower it? Will it work for her third period? She revises until it meets her standard. The user gets a quiz. The director gets an assessment designed to tell her something.

Those are teaching decisions, not technical ones. They draw on what teachers already know about content, students and instruction, which is the heart of Mishra and Koehler's (2006) TPACK framework – technology improves teaching only when it's combined with pedagogical and content knowledge. They also draw on the same habits behind good questioning, a long tradition that runs from Socrates through Bloom's taxonomy (Bloom, 1956; Anderson and Krathwohl, 2001) to the Question Formulation Technique (Rothstein and Santana, 2011). Be clear about purpose. Pay attention to context and the level of challenge. Evaluate the result against criteria.

Directing well also means knowing when not to trust the assistant. The research on automation is consistent about one hazard: people who rely on capable automated systems tend to watch them less closely and accept their errors, which is usually called automation bias or complacency (Parasuraman and Manzey, 2010). The consultant study in Part III found the same thing with generative AI. Over time, a tool that does the preparation well could erode the very judgment it depends on. The framework pushes back on that in a few ways. Build Day includes practice catching errors and lost rigor. Coaching focuses on the calls teachers make about AI output. And the quarterly work-sample review checks whether judgment is growing or slipping. Is that enough? I don't know yet, and it's one of the questions the evaluation is set up to answer.

So the professional development in this framework isn't a course in using AI. It's PD in the instructional strategies the diagnostic calls for, delivered through a tool that makes those strategies practical, with coaching aimed at the decisions teachers make about what their students need. Almost everyone starts out as a user of a new tool. Over the course of a year, the work is meant to help teachers become directors.

Part III: The Evidence Base

None of the pieces of this framework are new. Each one has its own body of research behind it, and what's changed is that the technology now makes it practical to put them together. In this section I lay out what the research says, how much weight it can carry, where it runs out and where it points in the other direction.

Teachers Matter Most, and the Workforce Keeps Turning Over

If there's one finding education research keeps coming back to, it's that teachers are the most important in-school factor in how much students learn. The Tennessee STAR experiment randomly assigned students and teachers to classrooms, and later analyses found large differences in achievement depending on which teacher a student happened to draw (Nye, Konstantopoulos, and Hedges, 2004). Chetty, Friedman and Rockoff (2014) followed the effects of teacher quality all the way out to students' earnings as adults.

And a district's teaching force never sits still. When I managed the Schools and Staffing Survey, our follow-up survey found that about 8 percent of public school teachers moved to a different school in a single year, and a similar share left teaching altogether. Turnover was concentrated among newer teachers and in high-need schools (Luekens, Lyter, and Fox, 2004). Ten years on, not much had changed – in 2012-13, about 8 percent of public school teachers moved and 8 percent left (Goldring, Taie, and Riddles, 2014). So any strategy built on teacher quality has to invest in the teachers who stay. If AI takes some of the after-hours load off that drives burnout, it could help with retention, too. That's a reasonable expectation. It isn't a finding yet.

What Works, and by How Much

Researchers broadly agree on which instructional practices are associated with the largest gains. The estimates below, though, come from very different kinds of studies and aren't on the same scale, and that matters when you read them. I've ordered them by the strength of the research design, with causal and randomized estimates at the top and broad syntheses at the bottom. When it comes to choosing specific practices, the prescription leans mainly on the What Works Clearinghouse practice guides and the Education Endowment Foundation's Teaching and Learning Toolkit. Both grade the evidence behind each recommendation, and both line up with the evidence language in federal law.

Practice or program	Estimated effect	Type of evidence
Instructional coaching	0.49 (practice); 0.18 (achievement)	Meta-analysis of causal studies (Kraft et al.)
Curriculum-focused professional development (STEM)	0.21 average across programs	Meta-analysis of experimental and quasi-experimental studies (Lynch et al.)

AI tutoring, student-facing (Khanmigo)	0.06–0.08 per term; ≈ 0.14 per year	Randomized trial (Oreopoulos and Low)
Formative assessment	$d \approx 0.20$	Meta-analysis (Kingston and Nash) [†]
Metacognitive strategies	+7 months' progress	Evidence review (EEF)
Higher-cognitive questioning	$d = 0.73$	Meta-analysis (Redfield and Rousseau) ^{**}
Classroom discussion	$d = 0.82$	Meta-synthesis (Hattie) [*]
Scaffolding	$d = 0.82$	Meta-synthesis (Hattie) [*]
Teacher clarity	$d = 0.75$	Meta-synthesis (Hattie) [*]
Feedback	$d = 0.73$	Meta-synthesis (Hattie) [*]
Explicit (direct) instruction	$d = 0.59$	Meta-synthesis (Hattie) [*]

[†]Black and Wiliam's (1998) review established formative assessment as a high-leverage practice. Later work argued that even Kingston and Nash's more modest estimate may be inflated by weak study designs (Briggs et al., 2012).

^{**}Samson, Strykowski, Weinstein and Walberg (1987) reanalyzed the questioning research and found considerably smaller effects; the direction of the finding has held up better than its size.

^{*}Values from Hattie (2009) and later Visible Learning updates. Hattie's method pools meta-analyses of uneven quality and has well-documented limitations (Snook et al., 2009; Slavin, 2018). The figures are useful for comparing practices with one another, not as predictions of what a program will produce.

Anyone who has spent time with assessment data will spot the scale problem right away. Effect sizes from broad syntheses run far higher than what randomized trials of real programs in real schools turn up. Kraft (2020) gives a more realistic yardstick for causal studies of achievement, where below 0.05 is small, 0.05 to 0.20 is medium and above 0.20 is large. Measured that way, coaching's 0.18 and curriculum-focused professional development's 0.21 sit near the top of what school programs produce. Khanmigo's 0.14 for a year of low-cost use is a respectable result too. Those causal and randomized numbers are the benchmarks this framework should be held to.

AI on Its Own Versus AI in Skilled Hands

In 2026, Oreopoulos and Low published a two-year randomized trial of Khanmigo, an AI tutor, in 18 Tennessee middle schools. Students gained 0.06 to 0.08 standard deviations per term in math, or about 0.14 over a full year of active use. Almost every student tried it. Most, however, used it in only about a third of the sessions available to them. The tool itself held up reasonably well; what limited the results was whether students engaged with it, and how.

The research on AI that supports adults is closer to what this framework does. In a randomized trial with about 900 tutors and 1,800 students, Tutor CoPilot fed human tutors real-time AI suggestions during their sessions. Students whose tutors had access were 4 percentage points more likely to master the material, and 9 points more likely when their tutor had been rated less

skilled. Tutors using the tool asked more guiding questions and gave away fewer answers (Wang et al., 2024). A smaller quasi-experimental study in one middle school found that students grew more in math when human tutors worked alongside AI tutoring software than when students used the software alone (Gurung et al., 2026). That design can't rule out other explanations. Still, the direction is the same as in the randomized work.

There's now research on teachers directly. In a randomized trial run by the Education Endowment Foundation (2024), secondary science teachers in England who used ChatGPT with structured guidance cut their lesson-planning time by about 31 percent, and an expert panel found no difference in quality. In the Gallup-Walton survey (2025), teachers who use AI weekly report saving about 5.9 hours a week. Those are self-reports from teachers who chose to use AI, so I treat that number as a ceiling. A third study, also randomized, gave instructors automated feedback on their classroom talk, and they went on to take up and build on students' ideas more often (Demszky et al., 2024). The instructors were volunteers teaching a large online programming course – a long way from a K-12 classroom – but it's an early sign that AI can support the teaching itself and not only the preparation.

The clearest warning comes from a randomized trial with about 1,000 high school math students in Turkey. Students who had open access to GPT-4 during practice did 48 percent better on practice problems, and then 17 percent worse on a later exam taken without it. A version built as a tutor, one that guided students toward answers rather than handing them over, improved practice performance by 127 percent and largely avoided the damage on the exam (Bastani et al., 2025). Yes, the study is about students, not teachers. The lesson carries over anyway. AI can lift immediate performance while weakening the underlying skill, and the design decides which of those you get.

Other professions show the same pattern. Customer-support agents with an AI assistant resolved 14 percent more issues per hour, and the least experienced agents improved by 34 percent (Brynjolfsson, Li, and Raymond, 2025). Consultants using AI on tasks within its capabilities finished 12 percent more tasks, 25 percent faster and at noticeably higher quality. On a task just beyond those capabilities, though, they were 19 percentage points less likely to get the right answer than colleagues working without it (Dell'Acqua et al., 2023).

Taken together, these studies point in one direction. When people are handed AI without structure, the gains are modest and there can be real harm. The larger gains show up when AI is built for specific work and used by people who exercise judgment about when to rely on it. Most teachers using AI today have never had any support in developing that kind of judgment.

The Foundations

Improvement science

Anthony Bryk and colleagues (2015) gave a name to a habit anyone who has served on a school committee will recognize – "solutionitis," or adopting a solution before anyone understands the problem. What you end up with, they wrote, is a system that "goes fast and learns slow." Their alternative is improvement science. It's adapted from Deming's quality-improvement methods, proven in healthcare, and built on disciplined Plan-Do-Study-Act cycles organized around a plain question. What exactly isn't working, and for whom? The Diagnose-Respond-Measure-Adjust cycle in this framework applies the same method. It adds a peer comparison as the starting point, and it uses AI to make acting on the data cheaper and faster.

Evidence-based instructional practice

Hattie's synthesis, whatever its flaws, gave educators a shared vocabulary for comparing practices. Kraft and colleagues supplied the strongest causal evidence for coaching. Lynch and colleagues showed what professional development built around curriculum can do, and the What Works Clearinghouse and the Education Endowment Foundation translated the research into graded guidance that schools can actually use. That's a sound enough basis for deciding where to put the money. In my experience the hard part has rarely been knowing what works – it's been making it practical for every teacher, in every class.

Professional judgment and transfer

What I'm calling the habits of a director (being explicit about purpose, about the students, about the level of challenge and about what good work looks like) are really just the habits of good lesson and assessment design. Some research on metacognition suggests habits like these carry over from one setting to another (Schraw, 1998). Willingham (2007) offers a fair counterpoint. Thinking skills are tightly bound to content knowledge, and they transfer less readily than we'd like. And the automation research raises a separate worry, since practicing with a capable tool can weaken judgment if the tool ends up doing the thinking. So my claim here is a narrow one. Practicing these judgments with the assistant, in the teacher's own content and with coaching, may strengthen them in the rest of her teaching. Which effect wins out is an empirical question, and this work will test it.

How organizations adopt AI

The workplace studies above share something. The biggest gains came where AI was built for a specific job and used inside the range of tasks it handles well, as with the customer-support assistant. The losses came where people leaned on it past those limits, as in the consultant study. The planning assistant takes that same approach to teaching. It's configured for specific work, loaded with the relevant knowledge, built into routines teachers already have, and used by

people who've been coached on when to trust it. Bryk made the case for improvement science partly on its track record in healthcare, and the same cross-sector reasoning applies here – although schools still have to test whether those lessons hold up in classrooms.

The Theory of Change

Here is the logic of the framework as a chain of claims, each one tied to research.

If a district starts with a diagnostic that pinpoints its gaps by grade, subject, school and student group (improvement science: make the problem specific),

then professional development can concentrate on the instructional practices that best match those gaps, organized around the district's adopted curriculum (Lynch et al.: professional development built around curriculum materials has larger effects; Kraft et al.: coaching improves both practice and achievement).

If teachers have a prebuilt system of agents prepared with their curriculum, standards and prescription, the assistant's output is accurate and grade-level, and teachers are coached in the prescribed practices and in the judgment needed to direct the assistant (TPACK; evidence on purpose-built AI in other fields),

then the preparation burden that has kept those practices from happening consistently will come down, and teachers will put at least part of the recovered time back into them.

If implementation and results are measured continuously, broken out by student group and adjusted each quarter (improvement science: you can't improve at scale what you don't measure),

then achievement in the targeted areas should rise more than it would with AI tools deployed without an instructional purpose. The benchmarks to beat are the Khanmigo trial, at about 0.14 standard deviations for a year of use, and coaching, at about 0.18.

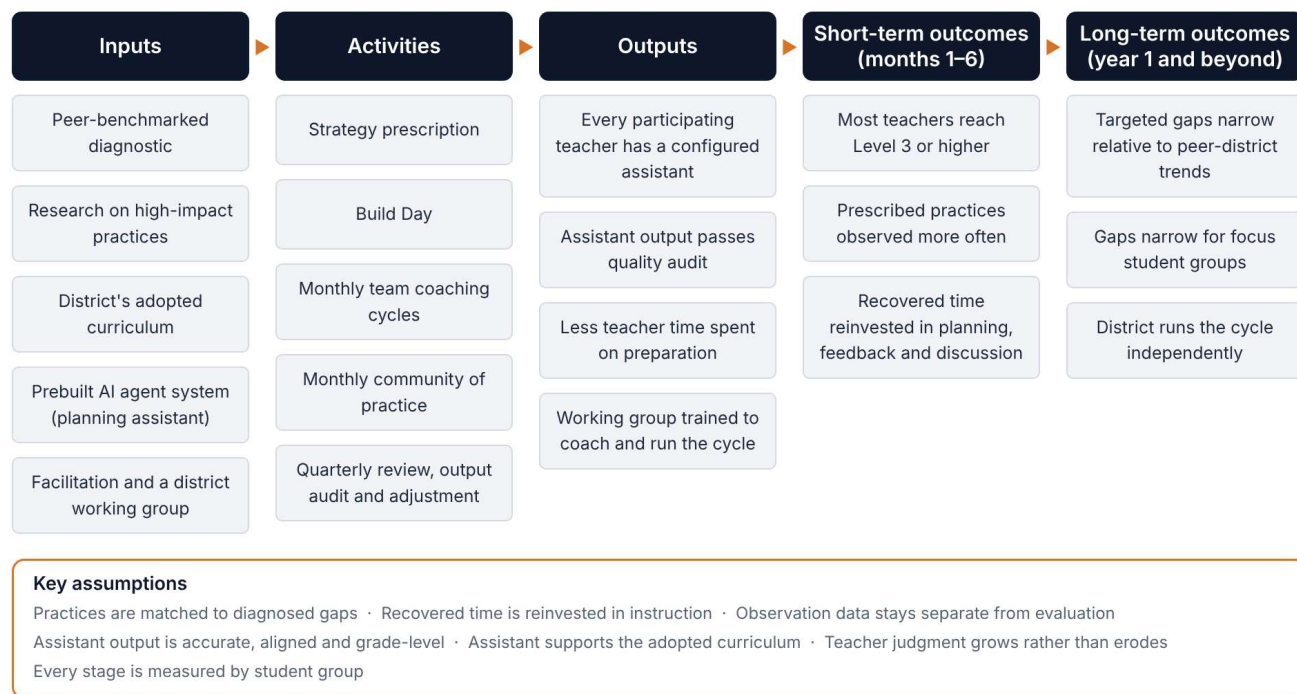


Figure 1. Logic model

Each link in that chain has research behind it, but three of them are weaker than the others. The first is the assumption that teachers will put recovered time back into instruction instead of losing it to everything else competing for it. I'm not willing to assume time savings take care of themselves, so the framework measures time use and classroom practice directly. The second is that the assistant's output is consistently accurate, aligned and rigorous; a sample of it gets audited every quarter for that reason. The third is that working with the assistant builds teachers' judgment and doesn't wear it down. Nobody has tested the chain as a whole system yet. The framework is set up as design-based implementation research (Penuel et al., 2011) for exactly that reason, and Part VI lays out the evaluation design.

How Strong Is the Evidence?

Superintendents and school committees deserve to know which parts of this stand on solid ground, which parts are still being tested and where the evidence cuts against it.

Well established

- Teachers are the most important in-school influence on student learning (Nye et al., 2004; Chetty et al., 2014).
- Coaching improves both teaching practice and student achievement, with smaller effects in larger programs (Kraft et al., 2018).
- Professional development built around curriculum materials is associated with larger effects than generic professional development (Lynch et al., 2019).

- Feedback, discussion, explicit instruction, metacognitive strategies and formative assessment have consistently positive evidence, though estimates of size vary widely by method (Hattie, 2009; EEF; Kingston and Nash, 2011).
- One-time workshop PD rarely changes practice (TNTP, 2015; GAO, 2026). Effective professional development is content-focused, collaborative, coached and sustained (Darling-Hammond et al., 2017).
- Most teachers have not received AI training from their school or district (Gallup-Walton, 2025).

Promising but early

- Student-facing AI tutoring produces small-to-medium gains, held back by low engagement (Oreopoulos and Low, 2026).
- AI that supports a human tutor improves student mastery, especially for less experienced tutors (Wang et al., 2024; Gurung et al., 2026).
- AI-assisted lesson preparation cuts planning time by roughly a third without lowering quality (EEF, 2024).
- In other professions, AI raises productivity most for less experienced workers but reduces accuracy on tasks beyond its capabilities unless people judge well when to rely on it (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023).
- Schools serving more disadvantaged students are integrating AI more slowly (Campos and Singleton, 2026).

Evidence that cuts the other way

- Open access to AI can raise immediate performance while weakening learning (Bastani et al., 2025).
- People who rely on capable automation tend to miss its errors (Parasuraman and Manzey, 2010; Dell'Acqua et al., 2023).
- Teacher-sourced supplementary materials are often low in quality (Polikoff and Dean, 2019).
- Broad syntheses overstate effects relative to causal studies (Snook et al., 2009; Slavin, 2018; Kraft, 2020).
- The coaching programs behind the 0.18 estimate generally involved more contact than this framework's monthly team cycles.

What this work will test

- Whether teachers reinvest the time AI saves in high-impact instructional practice.
- Whether the assistant's output is consistently accurate, aligned and rigorous.
- Whether team coaching supported by a prepared assistant can produce effects approaching those of more intensive coaching.
- Whether the full system improves achievement more than AI tools alone.

- Whether director judgment developed with the assistant grows, rather than erodes, and shows up in teachers' questioning and assessment design more broadly.
- Whether a district can run the cycle independently after one year.

Under the Every Student Succeeds Act, the approach meets Tier 4 evidence today, "demonstrates a rationale," which means a well-specified logic model grounded in research, with an evaluation planned. The evaluation in Part VI is built to move it higher. Correlational analysis with statistical controls for selection would meet Tier 3. A well-run quasi-experimental comparison with demonstrated baseline equivalence could meet Tier 2, and a randomized staggered start across a large enough, multi-site sample could meet Tier 1 (U.S. Department of Education, 2016).

Part IV: The Framework

Diagnose, Respond, Measure, Adjust

The framework applies improvement science to AI-integrated instruction. Each phase works inside the district's existing structures, curriculum and staff. And each one keeps coming back to the same question. Which students are being well served, and which ones aren't?



Figure 2. The improvement cycle

Diagnose: where are we, and who is being well served?

A recent diagnostic for a suburban Massachusetts district shows why the peer comparison matters. By the usual measures the district looked healthy: graduation rates above the state average, a stable teaching staff, engaged families. Compared with demographically similar districts, though – matched on enrollment, economic profile, English learner population and community characteristics – it was at or above the state average in 18 of 19 tested grade-subject areas, yet below its peer average in 16 of them, by 7 to 28 percentage points. Grade 3 was a genuine bright spot, above its peers in both ELA and math. But Grade 4 ELA proficiency had fallen to 36 percent while peers averaged 62, and Grade 8 math stood at 42 percent against a peer average of 70, with a median Student Growth Percentile of 36. Those gaps were systemic, and none of them showed up in the district's own trend lines.

The diagnostic happens in two stages.

- **The diagnostic snapshot** is free to any Massachusetts district and is built entirely from public DESE data, drawing on nine years of MCAS results across 321 Massachusetts districts. It picks five peer districts using six measures (enrollment, low-income percentage, English learner percentage, high-needs percentage, per-pupil spending and student-teacher ratio), combined in a weighted similarity score. The method is transparent and reproducible, and districts are welcome to suggest different peers. The snapshot then compares the district with those peers, and with the state, on MCAS achievement and Student Growth Percentiles by grade, subject, school and student group. It tracks trends back to before the pandemic and ends with a short list of the biggest opportunities and the strengths worth building on. Districts get it within a few days, as a report a leadership team or school committee can use.
- **The full diagnostic** is the first phase of an engagement, and it adds what public data can't show. That includes results by standards cluster, the district's interim assessments (MAP Growth, i-Ready, DIBELS or district benchmarks) read alongside MCAS, attendance and staffing context, and how the adopted curriculum is actually being used. It also takes an honest look at current practice – where the strongest classrooms are and what those teachers are doing, how teachers already use AI, and what past PD has produced. And it follows each class of students from one year to the next, so a strong or weak cohort doesn't get mistaken for a problem at a particular grade.

A few disciplines hold throughout. Results for small groups get reported only when the numbers are large enough to mean something, with multi-year averages where needed. Gaps are identified from multi-year averages too, not a single year's results, because a gap that gets picked for being unusually large in one year will tend to shrink on its own the next. Strengths get the same careful attention as gaps. The classrooms already beating the peer average are often the best source of ideas for what to do next. DESE's District Analysis and Review Tools (DART) already identify comparable districts, and the diagnostic builds on that idea with finer-grained comparisons and a direct line from each gap to the curriculum units and practices that address it.

Respond: a prescription, not a menu

I've seen too many diagnostic reports end up on a shelf because they stop at the findings. This one ends in a strategy prescription, made up of three to five prioritized instructional practices matched to the district's specific gaps. Each one is tied to how the planning assistant will be configured, what coaching will focus on and what will be measured.

Three questions drive the prescription. **Where are the largest gaps relative to similar districts, and for which students?** The answer sets the priority grades, subjects and student groups. **What kind of gap is it?** Weakness in analysis and synthesis in ELA calls for structured discussion and better questioning. Missing foundational math skills call for scaffolding and tight feedback loops. High absenteeism and wide variation between classrooms call for a different

starting point entirely. **What can the district take on now?** A district with a strong collaborative culture and common planning time can move into peer coaching quickly, while a district with high turnover needs the assistant to carry more of the preparation load at first.

One principle applies everywhere. Scaffolds are there to get students to grade-level work, and they shouldn't swap it out for easier work. TNTP's *The Opportunity Myth* (2018) found that students spend much of the school day on below-grade-level assignments, and that students of color and students from low-income families get them most often. A tool that makes differentiation easy could make that worse if all it did was lower the bar. So the assistant's default is to support access to the same rigorous task in the adopted curriculum.

For each practice, the prescription spells out the research behind it and its limits, how it connects to the adopted curriculum and how the assistant will be configured to support it. It also lays out what coaching will cover and in what order, how the practice will reach the students furthest behind, and which leading and lagging indicators will be reviewed each quarter.

Two districts can look alike on paper and still need very different prescriptions. Take a district whose Grade 6 students are well below peers in math, with the biggest gaps in ratios and proportional relationships and the widest gaps for economically disadvantaged students. Its prescription puts scaffolding and formative feedback in that domain first. The assistant gets error-analysis, misconception and scaffolded-practice routines built for proportional reasoning rather than generic math support. Coaching starts with formative data and scaffolding before it moves to discussion, and the measures track proportional reasoning for economically disadvantaged students specifically, not overall math scores. A second district might show solid foundational skills but flat writing, with the gap concentrated in analysis and argument. It gets structured discussion and rubric-aligned feedback on writing instead, along with a different configuration, a different coaching sequence and different measures.

In every case the entry point is the same: AI that saves teachers time. The destination is the instructional shifts that move student outcomes. The path between the two is specific to each district, and it's written down in the prescription.

Measure: is it working, and for whom?

Measurement starts in the first month, and it tracks what the diagnostic said needed to change. (Login counts tell a district very little.) There are five levels.

- **Implementation.** Are the prescribed practices showing up in the classrooms where the diagnostic said they were needed? Brief, focused observations by coaches and the working group answer that, along with samples of teachers' work with the assistant (shared with their consent) and short time-use logs that show where recovered time is

going. None of this information is used in educator evaluation. That's agreed with the local association before a single observation takes place.

- **Assistant quality.** Every quarter, a random sample of the assistant's output is rated for accuracy, alignment and grade-level rigor.
- **Leading indicators.** These are measures tied to the targeted standards and student groups. If the prescription targets proportional reasoning for economically disadvantaged students, then that's exactly what gets tracked.
- **Interim results.** The district's own interim assessments are reviewed each quarter against the diagnostic baseline, using a crosswalk to MCAS reporting categories built during the diagnostic.
- **Annual results.** MCAS achievement and growth, compared against the same peer districts used in the diagnostic. Did the gap close?

Every indicator is broken out by race, income, English learner status and disability. An initiative that raises averages but leaves the gaps where they were isn't succeeding. Subgroups at a single grade can be small, so quarterly results for small groups are treated as signals to look into rather than conclusions.

All of these measures serve the district's own improvement cycle. Figuring out whether the framework itself made the difference takes a comparison across schools and districts, which I describe in Part VI.

Each quarterly review should tell a simple story. The district invested in these practices because of these findings. Here's what changed in classrooms, and here's what the interim data shows about the targeted gaps, and for whom. And here's what we recommend next. I've sat through enough board presentations to know how rare that is.

Adjust: what do we change?

A quarterly review should end with decisions. Say a practice is working for most students but not for English learners. Then the assistant's routines pick up scaffolds for each proficiency level, the next coaching cycle and community session focus on multilingual learners, and an English-learner indicator goes into the next review. If classroom practice has clearly changed but results haven't, the prescription itself gets questioned – maybe the practice doesn't match the real gap, maybe the assistant needs reconfiguring, or maybe the measures are missing the signal. If the output audit keeps finding the same errors, fixing the assistant comes before anything else. And if results are improving but implementation is uneven, the answer is targeted support for the teachers and schools that need it. Every adjustment feeds into the prescription, the assistant, the coaching focus and the next quarter's targets. Because the cycle runs all year, the district ends up with a system that improves instead of a program that reports.

The system in practice

Here's an illustrative district. Its diagnostic shows Grade 4 ELA at 38 percent proficient against a peer average of 60. The gap is concentrated in the Integration of Knowledge and Ideas standards and is widest for students from low-income families. Grade 7 math sits at 45 percent against 66, concentrated in proportional reasoning and widest for students with disabilities. Both gaps have held for three years, so they aren't artifacts of one bad year.

The prescription calls for structured discussion and higher-order questioning in Grade 4 ELA, because the gap is in synthesis and analysis. In Grade 7 math it calls for scaffolding and formative feedback, with close attention to IEP accommodations. Grade 4 teachers' assistants come loaded with discussion protocols built around the targeted standards and the texts in the adopted units, along with text-based question sequences that move from literal to evaluative and quick checks on whether students can connect ideas across texts. Grade 7 teachers' assistants come with scaffolded practice on proportional reasoning tied to the adopted lessons, error-analysis routines that pick out specific misconceptions, and supports aligned with common IEP goals.

Build Day follows the prescription, and so does coaching. In the second month, Grade 4 teams work on leading text-based discussion and Grade 7 teams work on regrouping students by misconception. By the first quarterly review, observations show more frequent and stronger discussion in Grade 4, and the winter interim assessment shows gains on the targeted standards. But students from low-income families gained only about a third as much as their classmates. With roughly forty students in that group, one quarter doesn't prove anything. It does match what teachers and coaches are seeing, though – some students aren't getting into the discussions. So Grade 4 adds vocabulary and background-knowledge work ahead of discussions, the assistant's routines are updated to include it, and the next coaching cycle focuses on it. In Grade 7, what's working keeps going and gets extended to other strands with similar gaps.

Part V: How It Works in Schools

Relief First

Most PD ignores a simple rule, and this framework is built on it: teachers have to walk out with something they'll use the next morning. Every minute of professional development comes out of a day that's already full. So the first experience has to give time back – differentiated materials in a few minutes, exit tickets sorted before the next class, the parent email that otherwise gets written at 9 PM. If PD adds to a teacher's workload before it lightens it, it will struggle no matter how good the content is. Trust comes from relief, not from promises. None of this is new, either. Adults commit to learning that solves a problem they have right now and that they can put to use right away (Knowles, 1984).

The AI Planning Assistant

The planning assistant is the agentic harness from Part II. Think of how a good school gets a new colleague ready: you brief them on the curriculum, the standards and the school's priorities before day one. The assistant gets the same treatment. It helps with preparation and paperwork, and that's all it's for. It doesn't work with students, and it doesn't take over anything paraprofessionals or other staff do. It's built in four layers. The first three, which make up roughly 80 percent of the configuration, are finished before a teacher ever sees it.

- **Layer 1: Foundation.** The Massachusetts Curriculum Frameworks and the structure of MCAS; the library of prebuilt agents for instructional and administrative work (lesson preparation, differentiation, formative analysis, discussion and questioning, feedback, family communication, IEP progress-note drafts); and the instructional principles behind every response, including the default toward grade-level work. By default, differentiated material comes in three versions – approaching, on-grade and extending – all aimed at the same grade-level task.
- **Layer 2: Curriculum.** The scope and sequence of the district's adopted programs (enVision Mathematics, myView Literacy, i-Ready or others), along with unit maps, standards crosswalks, vocabulary and common misconceptions. If a teacher says "I'm starting Unit 5 on Monday," the assistant already knows what Unit 5 covers and what comes before and after it. It supports the adopted lessons; it doesn't replace them. I write these maps once per program and reuse them across districts, and publisher materials themselves are only used as far as the district's license allows.
- **Layer 3: District.** The diagnostic findings, the strategy prescription and how the two connect. With that in place, the assistant knows which units carry the standards where the district's students are furthest behind, and it shifts its defaults to match.

- **Layer 4: Teacher.** Her classes, her preferences and where she is in the curriculum, all set up on Build Day. This is the last 20 percent, and it's what makes the assistant hers. Students are identified by initials or a number. Never by name.

Because three of the four layers are in place ahead of time, the assistant is designed to beat a generic tool from the very first request – even a vague one. Type "quiz on Chapter 3" and you should get a quiz aligned to the right standards, with differentiation built in. That early quality matters. It's what earns a teacher's trust and keeps her using the tool long enough to get better with it. Whether the quality actually holds up is something the quarterly output audit checks.

Here's what day one is designed to look like. A teacher asks for practice on equivalent fractions for tomorrow. A couple of minutes later she has three versions of the grade-level task from her unit, each with a different level of support, plus an answer key. She pastes in last week's exit-ticket results and gets her students grouped by standard, with suggestions for reteaching. She asks for tomorrow's lesson and gets a warm-up, discussion questions that move from literal to analytical, practice, and an exit ticket, all fitted to the adopted lesson. Then she knocks out the parent email she's been putting off, in about ninety seconds. And a teacher with Algebra I, Geometry and AP Calculus? One assistant that knows all three.

The assistant isn't a separate product. It's built inside whatever AI platform the district already uses or chooses (Claude, ChatGPT, Gemini, or an education platform such as MagicSchool) as a configured system of agents with its own instructions, knowledge files and workflows. I build the first version alongside the district's working group, so the people who will maintain it understand how it works from the beginning. The platforms don't all support the same things. On some, the harness is a set of configured projects or custom assistants; on others it can include multi-step workflows. Districts hear which capabilities are available on each platform before they pick one, and Appendix C describes what actually gets built. The configuration belongs to the district and is documented so it can move to another platform, although anything platform-specific has to be rebuilt. The free educator versions of the major platforms run through June 2027. After that, districts will need education or enterprise licenses, and I help them choose and budget for those. Only accounts covered by the district's data privacy agreements are used.

Build Day

Build Day is one full professional development day. It isn't a workshop on how to use AI. I designed it around nine goals: trust, professional identity, immediate relief, capability, ownership, understanding, connection, vision and confidence. I lead it, with members of the working group co-facilitating, so teachers see from the first day who will be carrying the work in their schools. Larger districts can split it into more than one session.

We start the morning with people, not technology – introductions, why this work matters, and the idea that teachers' expertise is the scarce resource here and AI is what amplifies it. Then a

live demonstration. No slides; real output, generated in real time, with examples pulled from the district's adopted units and its diagnostic priorities. After that, teachers sit down with their own assistants and work through five real tasks using their own content. The first one is whatever administrative task they most want off their plate. Relief comes first. Capability follows.

The afternoon is more hands-on. Teachers personalize their assistants and test them on their hardest preps, and they practice catching errors, mismatched standards and lost rigor. Then they break into department groups to work on their own units. I hold the diagnostic findings until after all of this, when the data explains things rather than threatens people. We close by forming the district's working group, launching a 30-day challenge and taking a baseline survey. Teachers go home with materials they'll use the following week, five or more routines they can repeat on their own, and the start of the judgment they'll build over the year. Appendix B has a sample agenda.

From User to Director

Over the year, I expect teachers to move through five levels. The levels are about instructional decision-making rather than technical skill, and they're never used in evaluation. The timelines attached to them are working hypotheses; the quarterly work-sample review will test them.

- **Level 1, Accepting.** Simple requests, and the output gets used more or less as delivered: "Give me a quiz on Chapter 3." The prepared assistant is meant to make even this level useful. Build Day starts here, and that's fine.
- **Level 2, Specifying.** Now the teacher brings her instructional intent – the class, the unit, the range of students, the level of thinking she's after. "I need a warm-up for Period 3 on slope-intercept form. Some of them still struggle with plotting points." I'd expect this within the first six weeks.
- **Level 3, Evaluating.** She judges the output against professional standards and pushes back on specifics. "This version lowers the rigor instead of scaffolding access." "The exit ticket is at DOK 1, but I taught to DOK 3." Expected somewhere in months two through four, and it's often sparked by seeing colleagues' work in coaching and community sessions.
- **Level 4, Anticipating.** Formative evidence drives what she does next: "Last week's exit ticket showed 40 percent of Period 3 hadn't mastered equivalent fractions. Build scaffolded practice on that before I move to Unit 4." Teachers at this level build their own recurring routines. Expected in months four through eight.
- **Level 5, Designing.** The assistant becomes a thinking partner for designing instruction within the curriculum. "I want a three-week inquiry sequence on evolution that moves from observation to hypothesis to evidence to argument." These are the teachers who become the internal leaders and keep the work going. Expected in months eight through twelve.

To track progress, teachers fill out a short self-assessment at baseline, 30 days, 90 days and year end. Each quarter the working group also reviews samples of their work with the assistant (shared with their consent) against a simple rubric in Appendix B, and a subset gets scored by two trained reviewers so agreement can be checked. Put that alongside observations and student data and you can see whether the chain actually holds – does judgment develop, does instruction change, do results improve?

Coaching and Community

Of all the options, coaching has the strongest causal evidence for changing what teachers do, so it's built into the framework. It's also the hardest piece to scale, and the design is shaped around that problem.

- **Monthly coaching cycles in small teams, led by the district's own coaches.** Coaching is organized by grade-level or department teams of four to six teachers who are all working on the same prescribed practice. Each cycle has a team planning conversation, brief observations of each teacher focused on that practice, and a team debrief, plus individual follow-up where it's needed. The district's instructional coaches and working-group members lead the cycles. I train them before the first cycle, model a first cycle with each of them, and then observe their coaching and give them feedback through the year.
- **A monthly community of practice,** about an hour long and held within existing PLC or faculty meeting time. Teachers share what's working, work through problems and see how their colleagues are approaching things. The sequence follows the prescription rather than a fixed curriculum. I co-facilitate the early sessions with the working group; they lead by month six and are running the work on their own by month twelve.

Coaching the coaches is what makes this sustainable. With 80 participating teachers, team cycles mean roughly 15 a month instead of 80 individual ones, and the district's coaches lead them. My job is to get those coaches good at the work and support them while they do it. When the engagement ends, the coaching capacity stays in the district, which is the point.

The trade-offs are real. Teachers get less individual attention than one-on-one coaching would give, and the quality of the coaching depends on how well district coaches learn the work – which is why training and observing them is where most of my time in a district goes. Kraft and colleagues found no clear relationship between coaching hours and effects, which suggests focus matters more than volume. Still, the programs in their review generally involved more contact than one cycle a month. So can a prepared assistant plus team coaching by district staff deliver some of what intensive coaching delivers, at a fraction of the cost? That's one of the central questions of this framework, and the evaluation is designed to answer it.

The time commitment is modest on purpose, and all of it uses teachers' own classes and materials. Districts can award Professional Development Points for participation.

What It Looks Like in Classrooms

- **Math.** A teacher's students are struggling with proportional reasoning. She has her assistant produce three real-world scenarios at different levels of complexity, built on the problems in her adopted unit, each with a deliberate error for students to find, and she builds a structured discussion around them. Since the logistics of differentiation are already handled, she can keep her attention on how her students are reasoning.
- **English Language Arts.** In a persuasive writing unit, a teacher uses her assistant to review drafts against her rubric and draft specific feedback for each student, along the lines of "Your third paragraph claims community support but gives no evidence. What data would strengthen it?" She reads and edits every comment before a student sees it. Along the way she picks up patterns across the class that shape her next mini-lesson. Giving feedback like that on every draft, for 25 or more students, is very hard to keep up without help.
- **Science.** Planning a force-and-motion lab, a teacher asks her assistant for inquiry prompts for each lab group – questions, not answers. "What variable are you testing? What are you holding constant? What result would disprove your hypothesis?" She puts them on cards and uses them in conversation, and spends her own time with the groups on experimental design.

Who Does What

Every engagement is designed so the district carries the work and my role is to build its capacity to do it. The district names an internal AI lead and a working group of teachers, instructional coaches and school leaders. Over the year they take on the coaching, the community sessions, the upkeep of the assistant's curriculum and district layers, and the Diagnose-Respond-Measure-Adjust cycle itself. That follows what implementation research has found for decades: new practices take hold when training is followed by coaching on the job and backed by organizational structures that keep them going, and they fade when outside experts deliver training and leave (Fixsen et al., 2005). The goal of any engagement is to make itself unnecessary. The table shows how the work is divided and how it shifts.

Work	DirectEd	District
Diagnostic and prescription	Builds the diagnostic; drafts the prescription with district leaders	Provides interim data and context; approves the prescription
Assistant configuration	Builds the first version of layers 1–3 with the working group; trains them to maintain it	Owns the configuration; maintains the curriculum and district layers by year end
Build Day	Designs and leads it, with working-group members co-facilitating	Provides the PD day; working group introduced to staff

Coaching cycles	Trains district coaches; models a first cycle with each; observes and gives feedback on their coaching	Instructional coaches and working-group members lead team cycles from month two
Community of practice	Supplies the sequence; co-facilitates early sessions	Working group leads by month six
Measurement	Supplies protocols, rubrics and templates; trains reviewers; analyzes data for the first quarterly reviews	Collects observations, work samples, time-use logs and output reviews
Quarterly reviews	Facilitates the first reviews with district leaders	Leads the reviews by the second half of the year
After year one	Optional updated diagnostic and support on request	Runs the cycle independently

Equity is part of all of it. The assistant supports multilingual learners at different proficiency levels, flags possible bias in its output for the teacher to review, and produces materials in students' home languages where that helps, with Universal Design for Learning principles built in from the start. And since teachers are developing judgment rather than skill with one particular platform, what they learn carries over when the tools change. They will change.

Responsible Use

- **Accuracy.** Teachers review everything the assistant produces before any student sees it. The assistant names the unit and standard it's drawing on and flags uncertainty. On Build Day, teachers practice catching invented facts, mismatched standards, arithmetic errors and tasks that have lost their rigor, and a sample of output is audited each quarter.
- **Curriculum fidelity.** The assistant supports the district's adopted curriculum and doesn't replace core lessons. From time to time, community sessions look at whether supplementary material is drifting away from it.
- **Bias.** The assistant flags possible bias in what it produces, and community sessions periodically review materials for stereotyped examples, names and cultural assumptions.
- **Transparency.** When AI helps produce materials or feedback that students see, teachers are open about it with students and families, in line with district policy.
- **Staff use policy.** Each district gets a template staff acceptable-use policy, aligned with DESE's AI guidance, to adapt as it sees fit.
- **Student use.** Whether and how students use AI directly is a separate decision for the district. This framework doesn't deploy any student-facing tools.

Part VI: Evaluation Design

If I'm going to ask districts to measure everything, the framework has to be held to the same standard. This part lays out how the framework itself will be evaluated – the questions, the comparisons, the measures, the threats to validity and what gets reported. The evaluation works on two levels. Within each engagement, the district collects the implementation and outcome measures described here, using tools and training I provide, and results are reported descriptively. Across engagements, those data are pooled for the confirmatory analysis, which follows the registered plan. Before the first engagement begins, I will register the design, measures and analysis plan with the Registry of Efficacy and Effectiveness Studies, which is maintained by the Society for Research on Educational Effectiveness. Results then get reported against a plan that was set in advance.

Research Questions

I'm designating one question in advance as the confirmatory test: **does achievement on the targeted standards, for students in participating grades and schools, improve more than it would have without the framework?** Everything else is exploratory. Those results will be reported, and labeled that way, but they won't be used to claim success.

- Are the prescribed practices showing up more often in participating classrooms, and with higher quality?
- How much preparation time do teachers get back, and how much of it goes into instruction?
- Is the assistant's output accurate, aligned and rigorous – and does that hold across subjects and grades?
- Does director-level judgment grow over the year? If it does, does it show up in teachers' work beyond the assistant?
- Do effects differ by student group, and are the targeted gaps narrowing?
- After one year, can a district's working group run the cycle on its own?

Comparison Designs

I use two designs, alone or together, depending on what a district can support.

The first is a staggered start. Some schools or grade-level teams begin in the first semester and the rest start later. If a district agrees, a lottery decides the order, and the delay becomes a randomized comparison for that first semester. If the order is set by readiness instead, the comparison is quasi-experimental, and I check and report baseline equivalence between the early and late starters.

The second is a comparative interrupted time series. Here the district's results in its targeted areas are compared with those of its peer districts over several years before and after implementation. The question it answers is whether the district's trend changed relative to the trend it was already on. That's a much better question than whether this year beat last year.

There's also a third comparison inside the district itself: the standards and grades that weren't targeted. If targeted standards improve and untargeted ones don't, that's evidence the framework made the difference, rather than some general improvement. Teachers share materials across grades and schools, so I expect some spillover. That biases the comparison toward finding no effect, and it gets monitored.

Regression to the Mean

The diagnostic picks a district's largest gaps relative to its peers on purpose, and that creates a predictable problem. Some portion of any unusually large gap comes from a bad year, a weak cohort or plain measurement noise, and that portion tends to shrink by itself. A simple before-and-after comparison would hand the framework credit for that shrinkage.

The design handles this three ways. Gaps are identified using multi-year averages rather than a single year. Results are compared with trends, not with one baseline year. And I apply the same gap-selection rule to the peer districts, which lets the evaluation estimate how much their largest gaps narrowed with no intervention at all. Only improvement beyond that gets attributed to the framework.

Statistical Power

No single district can answer the confirmatory question by itself. When whole schools start at different times, statistical power depends mostly on how many schools there are, not how many students. Take assumptions that are common for achievement outcomes: an intraclass correlation of about 0.15, with prior achievement explaining about half of the student-level variation and 70 percent of the school-level variation. Under those assumptions, a district with ten schools split evenly between early and late starts could only reliably detect effects of roughly 0.4 to 0.5 standard deviations. That's about three times the benchmarks this framework is trying to beat. Pool about 60 schools across participating districts and the minimum detectable effect comes down to roughly 0.17.

So I'll report single-district results descriptively, as evidence about implementation and direction. The confirmatory analysis pools participating districts once enough of them have taken part. Final power calculations, based on the schools that actually participate and on Massachusetts-specific parameters, are part of the registered plan.

Measures

- **Classroom practice.** Coaches and the working group do formative observations using the look-fors in Appendix B, and those stay non-evaluative. For the pooled evaluation, a subset of lessons is scored with an established instrument that has published reliability, such as CLASS (Pianta, La Paro, and Hamre, 2008) or the Mathematical Quality of Instruction (Hill et al., 2008). The raters are trained on the instrument and aren't coaching those teachers, and at least 20 percent of observations are double-scored so agreement can be checked.
- **Director-level judgment.** Trained working-group members score work samples against the Appendix B rubric, with a subset scored by two reviewers and their agreement reported. Samples are requested at fixed points rather than letting teachers choose them, which cuts down on people sharing only their best work. Since samples are shared by consent, consent rates are reported and compare consenting and non-consenting teachers on whatever characteristics are available.
- **Time use.** Self-reported time savings are a ceiling. They aren't a measurement. Teachers keep short time-use logs for one week at baseline, mid-year and year end, recording time spent on preparation, paperwork, instructional planning and feedback. Where platform usage data is available and permitted, it supplements the logs.
- **Assistant output quality.** Every quarter, two trained reviewers from the working group, usually teachers or coaches in the subject, rate a random sample of the assistant's output for accuracy, standards alignment, grade-level rigor, fit with the adopted curriculum and bias, using a common protocol. I review the results and spot-check the ratings. Error rates are reported by subject and grade.
- **Student outcomes.** Interim assessments and MCAS, analyzed as described above and broken out by student group. Small groups are reported only when the numbers support it.

Threats to Validity

- **Developer as evaluator.** I designed this framework and I deliver it. That's the biggest threat to the credibility of any results, and I won't pretend otherwise. The safeguards are built into the design. The confirmatory question, measures and analysis are registered before any data exist, so I can't choose the flattering result after the fact. All results are reported against that plan and published whatever they show, and they're labeled as developer-reported. Districts are free to bring in their own evaluator, and with a district's permission, de-identified data can be made available to outside researchers who want to check the analysis.
- **Selection.** The districts and teachers who choose to take part are probably more motivated than average. Results may not generalize to districts that wouldn't have volunteered, and the reports will say so.
- **Regression to the mean.** Handled as described above.

- **Novelty.** New tools and extra attention can push effort up for a while. Effects are tracked across the full year, and into a second year where possible.
- **Concurrent initiatives.** Districts rarely change one thing at a time, so other initiatives in participating schools get documented and can be taken into account.
- **Turnover.** Teachers who leave or join mid-year are tracked. Analyses report how many students had a participating teacher for the whole year.
- **Multiple comparisons.** With this many grades, subjects and student groups, some differences will show up purely by chance. The confirmatory outcome is fixed in advance, and other results carry an exploratory label.

Ethics and Data

Research involving teachers and students needs formal review. An independent institutional review board will review the evaluation, and it will also go through each district's own research-approval process. Teachers consent separately to sharing work samples and to research observations. Saying no has no effect on their participation in the professional development. The student data used in the evaluation is limited to de-identified records covered by the district's data privacy agreement. No district's data is published without its permission. I'll publish combined findings across districts, disappointing ones included, and districts are always free to bring in their own evaluator.

Cost-Effectiveness

An effect size doesn't mean much without a cost attached to it (Kraft, 2020). A program with a modest effect that's cheap can be a better investment than one with a bigger effect at many times the price. Every engagement reports its full cost using the standard ingredients method (Levin et al., 2018). That includes the engagement fee, AI platform licenses, teacher and working-group time, stipends, substitute coverage, district staff time and a district data lead's time during the first month, divided by participating teachers and students. Because the coaching is done by district staff, their time is counted as a cost even though they're already on the payroll.

The comparison that matters is with what districts already buy. TNTP estimated roughly \$18,000 per teacher per year in the districts it studied. That figure counts staff time and isn't directly comparable, but it gives a sense of how much is already being spent on development. Intensive one-on-one coaching needs a full-time coach for a limited caseload of teachers. This framework's fee is kept under \$100,000 per engagement, which works out to roughly \$1,000 to \$2,000 per participating teacher in an engagement with 50 to 100 teachers, before licenses and district time. If it produced even part of coaching's effect at that cost, it would compare favorably. Does it? That's the whole point of the evaluation, and the cost-effectiveness ratio gets reported right alongside the effect.

What Would Count as Failure

If a framework can't fail, it can't be tested. So the registered plan spells out ahead of time which results lead to which conclusions:

- No change in classroom practice means the implementation model has failed, regardless of what happens to test scores. The coaching and Build Day design would need to be revised.
- If the assistant's output fails the quality audit at meaningful rates, the harness gets fixed before anything else is concluded.
- Practice changes but targeted achievement doesn't move? Then either the link between these practices and these gaps doesn't hold in this context, or the measures are missing the effect. Either way, the prescription logic has to be revisited.
- And if achievement improves, but by no more than in comparison schools or peer districts, the framework hasn't shown it made the difference.

Whichever of these happens, it will be published.

Part VII: Questions Districts Ask

I've sat on a school committee, so I have a good idea what gets asked when a proposal like this comes up in a public meeting. These are the questions I'd expect, and how I'd answer them.

Will this be used to cut staff?

No. Nothing in this framework reduces teaching or paraprofessional positions, and staff cuts play no role in its model or its economics. The time the assistant saves is supposed to go back into teaching – and the measurement system checks whether it actually does. Every engagement's terms, including the commitment to keep all observation data out of evaluation, get worked out with the local association before any work starts.

What does this ask of teachers?

One Build Day, held on a PD day the district already has. A monthly team coaching cycle, led by the district's own coaches, that runs about an hour to an hour and a half, during common planning time wherever we can manage it. A monthly community session of about an hour inside existing meeting time. And three one-week time-use logs over the course of the year.

Instructional coaches and working-group members take on more, and working-group members are typically paid stipends under existing agreements. What teachers get back is time, starting in the first week.

What will students notice?

Teaching that's more responsive and more demanding. If a student is behind, the support is aimed at a specific gap and at the same grade-level work everyone else is doing. A student who's ahead gets a task that actually stretches her thinking. A student with an IEP gets materials aligned with his real goals.

Whether students use AI tools themselves is a separate decision for the district. I'd point out, though, that teachers who direct AI well are far better prepared to teach the AI and digital literacy module the Graduation Council has recommended.

Will this replace our curriculum?

No. The assistant is built on the curriculum the district has already adopted, and its job is to support it – scaffolds, checks for understanding and discussion questions for the lessons teachers are already teaching. It doesn't write replacement units. Part of each quarterly review looks at whether supplementary material is drifting away from the curriculum. If your district has invested in high-quality instructional materials, think of this as a way to get more out of that investment.

Is student data safe?

The assistant is built from curriculum, standards and district-level results. Student records aren't part of it. Students are identified by initials or a number – never by name, student ID or any other identifying record – and anything involving individual records (finalizing IEP progress reports or grades, for example) stays in the district's own systems.

We only use education or enterprise accounts covered by a signed Massachusetts Student Data Privacy Agreement, consistent with FERPA, 603 CMR 23.00 and DESE's AI guidance. Vendors under those agreements may not use student data to train their models. And the diagnostic snapshot uses public DESE data only.

What will I be able to tell my school committee?

When I was on the committee side of the table, budget season came down to one question: what did we get for it? The diagnostic shows where the district stands against districts like it. The quarterly reviews then connect the investment to what changed in classrooms and to student results by student group, whether the news is good or not. At year end, the annual report adds the full cost per teacher and per student.

What does it cost, and what does it return?

Pricing comes separately. It's a single annual fee, scaled to the number of participating teachers and kept under \$100,000, and procurement details accompany each proposal. AI platform licenses after June 2027 are a separate district cost; the diagnostic includes a total-cost estimate so the committee sees the whole number.

The return gets reported every quarter, with a full cost-effectiveness analysis at year end (Part VI). The time savings in the research are real – about 31 percent less planning time in a randomized trial (EEF, 2024). But saved time only matters if it goes back into teaching, so we measure time use and classroom practice directly.

What happens after the first year?

By then the working group is running the coaching, the community sessions and the cycle. The assistants, and everything that configures them, belong to the district. They keep working as long as the district maintains its platform licenses, with the curriculum and district layers refreshed as programs change and new MCAS results come in. Districts that continue get an updated diagnostic and a revised prescription. Districts that stop keep everything they built.

Why is this equity work?

Because the achievement gap is turning into an AI gap. Campos and Singleton (2026) surveyed principals nationally and found that schools serving more disadvantaged students score lower on an index of AI integration covering policies, training, guidance and tools. Each standard deviation of additional disadvantage is associated with a score 0.07 to 0.11 standard deviations lower.

Well-resourced districts will find their way. Others are less likely to, and we could end up with one more layer of inequality. So equity is built into the structure here. Every diagnostic is broken out by student group, and every prescription is checked for its effect on the students furthest behind. Each review asks "for whom?" before it asks "is it working?" – and the assistant defaults to grade-level access rather than a lower bar.

Haven't we heard this before?

Yes, and the skepticism is earned. Interactive whiteboards were going to transform teaching. So were tablets, and so were personalized learning platforms. Larry Cuban documented the pattern in *Oversold and Underused* (2001): adoption, disappointment, repeat. It's the same churn Hess described in *Spinning Wheels*: one reform after another adopted and abandoned before anyone learns whether it worked. Three things are different this time.

First, the technology isn't the active ingredient. My own research taught me that structure alone rarely explains results. In a study of charter schools, autonomy by itself didn't predict teacher qualifications; what mattered was what schools did with it (Burian-Fitzgerald, Luekens, and Strizek, 2004). In the NYU study of New York City's small high schools, being smaller didn't make a school better on its own. What mattered was the "cohesion of pedagogical and non-academic practices" inside each school (Jacobowitz, Weinstein, Maguire, Luekens, and Fruchter, 2007). I expect AI to follow the same pattern. Giving teachers access to the tool is the structural change, and what a district builds around it will determine the result.

Second, measurement is built in from the start – a comparison group and a publicly registered plan, with results reported against that plan whatever they show. Part of the reason Cuban's cycle and Hess's churn keep repeating is that initiatives so rarely get measured before they're replaced.

Third, I'm going to be clear about what's proven and what isn't. The teaching practices are well established. Using AI to make them sustainable is the part being tested. Overpromising is how education technology lost teachers' trust in the first place, and I don't intend to do it again.

Part VIII: Working Together

This part covers how an engagement is structured. Commercial and procurement details come separately with each proposal. Every engagement starts with the free diagnostic snapshot described in Part IV. If the findings are useful, the district and I talk about what a full engagement might look like; if they aren't, the district keeps the snapshot anyway.

Engagement Timeline

There's no need to start everywhere at once. In fact, beginning with some grades or schools is often the smarter move. It keeps the working group's load manageable, and it creates the staggered comparison described in Part VI.

Timing	Phase	What happens
Months 0–1	Diagnose	Full diagnostic; prescription; district names AI lead and working group; coach training begins; measurement baseline and first time-use log; agreements with the local association; data privacy agreements confirmed; research approvals; evaluation plan registered
Month 2	Build Day	Assistants configured in advance with the working group; full day with all participating teachers, co-facilitated by the working group
Months 2–12	Coach and learn	Monthly team coaching cycles led by district coaches, with DirectEd training, modeling and feedback; community of practice co-facilitated, then led by the working group by month 6
Months 3, 6, 9, 12	Measure and adjust	Quarterly reviews and output audits, facilitated by DirectEd at first and led by the district from month 9; prescription, assistant configuration and coaching focus updated
Month 12	Transition	Annual report against peer districts, with full costs; working group runs the work; option for a second year

Targets & Checkpoints

Each district's targets are set during the diagnostic, based on its own baseline. The examples below give a sense of the commitments every engagement includes.

Timing	Illustrative target	If not met
Day 30	At least 75 percent of participating teachers using their assistant weekly; median time saved of two hours or more per week on time-use logs	Additional follow-up and workflow revisions

Month 6	At least 50 percent of teachers at Level 3 or higher on the work-sample rubric; prescribed practices present in at least half of observed lessons in focus grades; output audit error rates within agreed limits	Formal checkpoint to measure progress; transition to district-led CoP
Month 9	Interim gains on targeted standards for focus student groups at least as large as for other students	Prescription revised for those groups
Month 12	Targeted gaps narrowed relative to peer-district trends on MCAS achievement or growth; working group leading independently	Year-two decision based on the evidence

Implementation is a shared responsibility, so the targets are agreed jointly. I structure my work on the components that most directly support teachers and, in turn, student learning. Results depend on many factors, and they get reported in full either way. Everything built during the engagement stays with the district.

Capacity and Independence

I take on a limited number of engagements each year and am personally involved in each one. Because each district's own coaches and working group carry the day-to-day work, my time goes to the parts that build capacity: the diagnostic and prescription, the first configuration of the assistant, Build Day, training and supporting district coaches, and the first quarterly reviews. Every configuration, protocol and report is documented and belongs to the district, so the work doesn't depend on any one person, me included.

I don't receive payments from AI companies, curriculum publishers or platform providers, and I earn nothing from a district's choice of tools. I don't resell software or licenses. I follow the Massachusetts conflict-of-interest law (G.L. c. 268A) in all of my public-sector work and while I have previously held an elected role as school committee member in a district, I do not hold any public positions currently.

Building the Evidence Together

Each piece of this framework rests on something with evidence behind it. The new part is the combination. An honest peer comparison, and a prescription tied to it. AI prepared to make high-impact teaching practical inside the curriculum a district has already chosen. Coaching that builds teachers' judgment, and measurement that doesn't stop. The first districts to take part will help answer whether it works as designed. Their data will shape it, their students will benefit first, and what we learn together will help the districts that come after them.

School improvement gets built from the inside, by the people who do the work. I've seen what happens when it isn't – as a researcher, as someone who served on a school committee, and as the parent of two sons who came up through Massachusetts public schools. AI is going to

change schools whether districts plan for it or not. What worries me is that it ends up helping students in the districts that figured it out early, and leaves everyone else further behind.

Contact. Mike Luekens, DirectEd: mluekens@directed.pro | DirectEd.pro

About the Author

Mike Luekens' background brings together education research, large-scale assessment, executive leadership, data engineering, analytics, and AI.

At the American Institutes for Research, he managed the Schools and Staffing Survey for the National Center for Education Statistics, the most comprehensive national survey of K-12 schools and teachers, and authored federal reports including the NCES study of teacher attrition and mobility cited in this paper (Luekens, Lyter, and Fox, 2004). He also co-authored research on charter school autonomy and teacher qualifications (Burian-Fitzgerald, Luekens, and Strizek, 2004).

At Measured Progress, a longtime MCAS developer, he managed the full assessment cycle for the state of Utah, from development through scoring, scaling and reporting, for the state's high-stakes graduation test: the Utah Basic Skills Comprehensive Assessment program. The diagnostic in this framework draws directly on that experience with large-scale measurement.

He holds a B.A. in Sociology from Wake Forest University and an M.A. in Education Policy from George Washington University, where he completed the doctoral research methods and statistics sequence, and he completed Ph.D. coursework at New York University's Wagner School. As a research assistant at NYU's Institute for Education and Social Policy, he co-authored the study of small high school effectiveness cited here (Jacobowitz et al., 2007). He is a Fellow of the Institute for Educational Leadership.

He spent 17 years in senior leadership at Liberty Mutual, managing multi-million dollar budgets and teams of more than 350. He built NE Provenance, a property data platform covering 4.7 million New England properties, which supplies the data engineering behind the diagnostic. He also built QuestionCraft.ai, a platform for developing structured questioning and prompting skills; the framework doesn't require it, and districts aren't asked to use it.

As a former elected member of the Newburyport School Committee, he chaired the Subcommittee on Curriculum, Instruction, and Accountability. He has taught social studies to adults preparing for the GED at Academy of Hope in Washington, D.C. He is the father of two sons who have attended Massachusetts public schools.

References

- American Federation of Teachers. (2025, July 8). AFT to launch National Academy for AI Instruction with Microsoft, OpenAI, Anthropic and United Federation of Teachers [Press release].
- Anderson, L. W., and Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122, e2422633122.
- Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7-74.
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay.
- Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., and Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, 31(4), 13-17.
- Bryk, A. S., Gomez, L. M., Grunow, A., and LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.
- Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889-942.
- Burian-Fitzgerald, M., Luekens, M. T., and Strizek, G. A. (2004). Less red tape or more green teachers: Charter school autonomy and teacher qualifications. In K. Bulkley and P. Wohlstetter (Eds.), *Taking account of charter schools: What's happened and what's next?* (pp. 11-31). Teachers College Press.
- Campos, C., and Singleton, J. D. (2026). *AI diffusion gaps: Unequal integration of AI across K-12 schools* (NBER Working Paper 35347). National Bureau of Economic Research.
- Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633-2679.
- Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Harvard University Press.
- Darling-Hammond, L., Hyler, M. E., and Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute.
- Dell'Acqua, F., McFowland, E., Mollick, E., et al. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Working Paper 24-013). Harvard Business School.
- Demszky, D., Liu, J., Hill, H. C., Jurařsky, D., and Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3), 483-505.
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181-199.
- Doan, S., et al. (2024). *Findings from the 2024 State of the American Teacher Survey*. RAND Corporation.
- Education Endowment Foundation. (2021). *Metacognition and self-regulation*. Teaching and Learning Toolkit.
- Education Endowment Foundation. (2024). *ChatGPT in lesson preparation: A Teacher Choices trial*. Evaluation report.

- Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., and Wallace, F. (2005). *Implementation research: A synthesis of the literature*. University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.
- Gallup and Walton Family Foundation. (2025). *Teachers and AI: The AI dividend. How AI is helping teachers reclaim valuable time*.
- Goldring, R., Taie, S., and Riddles, M. (2014). *Teacher attrition and mobility: Results from the 2012-13 Teacher Follow-up Survey* (NCES 2014-077). U.S. Department of Education, National Center for Education Statistics.
- Gurung, A., Gao, G., Gutterman, J., Thomas, D. R., Gupta, S., Branstetter, L., Brunskill, E., Aleven, V., and Koedinger, K. R. (2026). Improving hybrid human-AI tutoring by differentiating human tutor roles based on student needs. In *Proceedings of the 19th International Conference on Educational Data Mining (EDM 2026)*.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hess, F. M. (1998). *Spinning wheels: The politics of urban school reform*. Brookings Institution Press.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., and Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.
- Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.
- Jacobowitz, R., Weinstein, M., Maguire, C., Luekens, M., and Fruchter, N. (2007). *The effectiveness of small high schools, 1994-95 to 2003-04*. Institute for Education and Social Policy, New York University.
- K-12 Statewide Graduation Council. (2026). *Final recommendations to modernize high school graduation requirements*. Commonwealth of Massachusetts.
- Kingston, N., and Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.
- Knowles, M. S. (1984). *Andragogy in action: Applying modern principles of adult learning*. Jossey-Bass.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241-253.
- Kraft, M. A., Blazar, D., and Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547-588.
- Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., and Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis* (3rd ed.). SAGE.
- Luekens, M. T., Lyter, D. M., and Fox, E. E. (2004). *Teacher attrition and mobility: Results from the Teacher Follow-up Survey, 2000-01* (NCES 2004-301). U.S. Department of Education, National Center for Education Statistics.
- Lynch, K., Hill, H. C., Gonzalez, K. E., and Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260-293.
- Massachusetts Department of Elementary and Secondary Education. (2025). *Massachusetts guidance for artificial intelligence in K-12 education*.
- Massachusetts Department of Elementary and Secondary Education. (2026). *2026 MCAS results*.

- Massachusetts Executive Office of Education. (2025, June 25). *Healey-Driscoll Administration launches "Future Ready: AI in the Classroom" for educators* [Press release].
- Mishra, P., and Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017-1054.
- National Center for Education Statistics. (2025). *The Nation's Report Card: 2024 reading and mathematics results*. U.S. Department of Education.
- Nye, B., Konstantopoulos, S., and Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237-257.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing.
- Oreopoulos, P., and Low, N. (2026). *One click away: AI tutoring with Khanmigo in a two-year school experiment* (NBER Working Paper 35620). National Bureau of Economic Research.
- Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.
- Penuel, W. R., Fishman, B. J., Cheng, B. H., and Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, 40(7), 331-337.
- Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) manual*. Paul H. Brookes.
- Playlab. (2026). *State AI guidance for K-12 education* [Policy tracker, accessed October 2026].
- Polikoff, M., and Dean, J. (2019). *The supplemental curriculum bazaar: Is what's online any good?* Thomas B. Fordham Institute.
- RAND Corporation. (2025). *American School District Panel: AI training and policies in K-12 districts*.
- Reason Foundation. (2025). *K-12 education spending spotlight*.
- Redfield, D. L., and Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research*, 51(2), 237-245.
- Rothstein, D., and Santana, L. (2011). *Make just one change: Teach students to ask their own questions*. Harvard Education Press.
- Samson, G. E., Strykowski, B., Weinstein, T., and Walberg, H. J. (1987). The effects of teacher questioning levels on student achievement: A quantitative synthesis. *Journal of Educational Research*, 80(5), 290-295.
- Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, 26, 113-125.
- Slavin, R. E. (2018, June 21). John Hattie is wrong. *Robert Slavin's Blog*.
- Snook, I., O'Neill, J., Clark, J., O'Neill, A.-M., and Openshaw, R. (2009). Invisible learnings? A commentary on John Hattie's book *Visible learning*. *New Zealand Journal of Educational Studies*, 44(1), 93-106.
- TNTP. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*.
- TNTP. (2018). *The opportunity myth: What students can show us about how school is letting them down, and how to fix it*.
- U.S. Department of Education. (2016). *Non-regulatory guidance: Using evidence to strengthen education investments*.
- U.S. Government Accountability Office. (2026). *K-12 education: Research on effectiveness of professional development is mixed, but teachers find collaborative learning most useful* (GAO-26-107874).
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., and Demszky, D. (2024). *Tutor CoPilot: A human-AI approach for scaling real-time expertise*. arXiv:2410.03017.

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, 31(2), 8-19.

Appendix A: Alignment with State and National Frameworks

AI. State and National Frameworks

The table below maps the framework to the standards and plans Massachusetts districts already work under. The point is for an engagement to fit inside commitments a district has already made, rather than show up as one more separate initiative.

Framework	Relevant elements	How this framework supports them
DESE District Standards and Indicators	Curriculum and Instruction; Assessment; Human Resources and Professional Development; Student Support	Diagnostic uses assessment data by student group; prescription and coaching target instruction; professional development is job-embedded and evaluated
High-quality instructional materials (DESE CURATE)	Adoption and strong implementation of highly rated core curricula	Assistant built on the district's adopted curriculum; supplements support core lessons rather than replace them; curriculum fit checked in output audits
Massachusetts educator evaluation framework	Standard I (Curriculum, Planning and Assessment); Standard II (Teaching All Students); Standard IV (Professional Culture)	Assistant supports planning, differentiation and formative assessment; coaching targets Standard II practices; working group builds collaborative culture. Observation data from this work is never used in evaluation.
Massachusetts Standards for Professional Development	Sustained, collaborative, job-embedded, data-driven, evaluated	Twelve months of team coaching and community, focused on the district's own data and curriculum and measured quarterly
Learning Forward Standards for Professional Learning (2022)	Evidence; Learning Designs; Implementation; Culture of Collaborative Inquiry; Equity Practices	Diagnostic and quarterly reviews; coaching cycles; working group; measures broken out by student group
DESE AI guidance (2025)	Training and tools; embedding AI in practice; privacy; human oversight	Curriculum-specific capacity building; data privacy agreements; teacher review of all AI output; quarterly output audits
Multi-Tiered System of Support	Data-based, tiered instruction	Formative analysis and grouping routines; scaffolds supporting Tier 1 access to grade-level work
Blueprint for English Learner Success; WIDA ELD Standards	Access to grade-level content with language support	Scaffolds by proficiency level; English-learner indicators in every quarterly review

IDEA and the IEP process; Universal Design for Learning	Instruction aligned with IEP goals; accessible design	IEP-aligned supports; UDL principles in the configuration; IEP documents finalized in district systems
K-12 Graduation Council requirements	Capstones; MyCAP; AI and digital literacy	Teacher fluency with AI; emphasis on higher-order, applied tasks
District and school improvement plans	Priority goals and measures	Prescription and targets built from the district's existing goals
ESSA evidence tiers	Tier 4 now; designs for Tiers 3, 2 and 1	Logic model; staggered and comparative designs; registered evaluation with results pooled across districts

A2. DESE AI Guidance

DESE guidance element	How this framework addresses it
Clear expectations for AI use	Built-in guardrails in the assistant's instructions; template staff acceptable-use policy
Student privacy	Students identified by initials or number only; signed Massachusetts Student Data Privacy Agreement; 603 CMR 23.00
Bias and misinformation	Assistant flags possible bias; teachers review all output; Build Day practice catching errors; quarterly output audits
Strengthen, not replace, instruction	AI handles preparation; the teacher handles instruction; measurement tracks whether instruction improves
Educator AI capacity	Twelve-month progression from accepting output to directing it with professional judgment

A3. Darling-Hammond, Hyler and Gardner (2017): Seven Features of Effective Professional Development

Feature	How this framework meets it
Content-focused	Tied to the adopted curriculum, specific standards and diagnosed gaps
Active learning	Build Day is hands-on from the start; coaching and community sessions work with teachers' own content
Collaboration	Team coaching cycles, monthly community of practice, working group, department blocks on Build Day
Models of effective practice	Live demonstration on Build Day; coaches model a first cycle; the assistant's output carries the prescribed practices
Coaching and expert support	Monthly team coaching cycles led by trained district coaches, supported by DirectEd

Feedback and reflection	Observation debriefs, work-sample reviews, director self-assessment and quarterly data reviews
Sustained duration	Twelve months of continuous work

A4. Desimone (2009): Core Features of Professional Development

Feature	How this framework meets it
Content focus	Curriculum-specific and driven by the diagnostic
Active learning	Teachers work with their own content rather than listening to presentations
Coherence	Aligned to the district's diagnostic, its adopted curriculum, state standards and DESE guidance
Duration	Twelve months, well beyond the 20 or more hours of contact time associated with change in the research Desimone reviews
Collective participation	Teachers from the same school, grade or subject work together; working group model

Appendix B: Implementation Toolkit

These are condensed versions of the tools I use in each engagement. Each one gets adapted to the district's prescription.

B1. Sample Build Day Agenda

Time	Segment	Goals served
8:00–8:30	Opening: introductions, why this work matters, teachers' expertise as the resource AI amplifies	Trust; professional identity
8:30–9:15	Live demonstration across subjects using district units and diagnostic priorities; no slides	Vision; understanding
9:15–10:30	Hands-on block 1: the administrative task each teacher most wants off their plate	Immediate relief
10:45–12:00	Hands-on block 2: this week's instructional materials for the teacher's own classes, built from the adopted curriculum	Capability
12:00–12:45	Lunch	
12:45–1:30	Personalization and stress test on the hardest prep; checking output for errors, misaligned standards and lost rigor	Ownership; confidence
1:30–2:15	Department deep dives; then the district's diagnostic findings	Understanding; connection
2:15–2:45	Working group formation; 30-day challenge; share-out	Connection; vision
2:45–3:00	Baseline survey and first time-use log instructions	Measurement

B2. Director-Level Rubric for Work Samples

With teachers' consent, a sample of their work with the assistant is collected at fixed points each quarter and reviewed against this rubric. A subset is scored by two trained reviewers to check agreement. Results are only ever reported in aggregate, and they're never used in evaluation.

Level	What the work sample shows	Typical evidence
1. Accepting	General requests; output used largely as delivered	Quiz or worksheet used without revision
2. Specifying	Instructional intent stated: class, unit, students, level of thinking	Range of learners described; standard or DOK named

3. Evaluating	Output judged against professional criteria and revised	Revisions that restore rigor, fix alignment or correct errors
4. Anticipating	Formative evidence drives the next instructional decision	Exit-ticket results used to plan the next lesson; reusable routines
5. Designing	Assistant used as a thinking partner for multi-week design	Unit or inquiry sequences co-designed and refined within the curriculum

B3. Classroom Look-Fors

Coaches and the working group use these look-fors for formative observations, which are non-evaluative. Each observation is brief (15 to 20 minutes), looks only at the prescribed practice, and gets shared with the teacher in the debrief. The pooled evaluation relies on an established observation instrument scored by trained raters who aren't coaching the teachers they observe (Part VI).

- **Structured discussion:** most students talk about the content, respond to one another, and get pressed by the teacher for evidence and reasoning.
- **Higher-order questioning:** a meaningful share of the questions call for analysis or evaluation, with adequate wait time and follow-up probes.
- **Scaffolding:** supports give students access to the grade-level task instead of replacing it, and they fade as students gain independence.
- **Formative feedback:** evidence of learning is gathered during the lesson, instruction adjusts in response, and feedback is specific and actionable.

B4. Quarterly Review Report Outline

- Summary and decisions requested
- Implementation: assistant use, distribution of levels, observed practice in focus grades
- Time use: where recovered time is going
- Assistant output quality: audit results and corrections made
- Leading indicators on targeted standards, by student group
- Interim results against baseline and comparison group
- Equity check: who it's working for, and who it isn't working for yet
- Recommended adjustments to the prescription, assistant configuration and coaching focus
- Status of current targets, and targets for the next quarter

Appendix C: Intervention Specification

This appendix describes the planning assistant in enough detail that another team could build it and evaluate it. I update it as the build evolves, and the version in use gets recorded with each engagement's evaluation.

C1. Components

Component	What it contains	Maintained by
Instructions	Role; instructional principles (grade-level default, scaffold to access, support the adopted curriculum, teacher reviews all output); output formats; requirement to name sources and flag uncertainty	DirectEd; reviewed annually
State knowledge	Massachusetts Curriculum Frameworks; MCAS structure and reporting categories	DirectEd; updated when frameworks change
Curriculum knowledge	Unit maps, standards crosswalks, vocabulary and common misconceptions for the adopted programs; publisher content only as licensed	DirectEd at first, then the working group
District knowledge	Diagnostic findings; prescription; priority units, standards and student groups	Working group; refreshed each quarter
Agents and workflows	Lesson preparation; differentiation; formative analysis; discussion and questioning; feedback; family communication and paperwork	DirectEd; revised after each quarterly review
Teacher layer	Classes, preferences and place in the curriculum; students identified by initials or number only	Teacher

C2. Platform Implementation

The harness uses each platform's native tools for persistent instructions, knowledge files and reusable workflows. Where a platform supports project-level instructions and knowledge, each agent is set up as a configured project or custom assistant. Where it supports multi-step workflows, agents can chain steps together – for example, finding the lesson in the adopted unit, drafting material aligned to its standard, and then checking that the scaffolds lead back to the grade-level task. If a platform supports less, the assistant does less, and districts hear which capabilities are available before they choose. Configurations are documented in plain files so they can move between platforms, although platform-specific features have to be rebuilt.

C3. Quality Checks

Each agent's instructions require it to name the unit and standard it's drawing on, flag uncertainty, check that scaffolds lead to the grade-level task, stay within the adopted lesson, and avoid inventing facts or sources. Those are instructions to the model, and the model doesn't always follow them. Language models still make mistakes. So teachers review every output before students see it, and a sample of output is audited each quarter (Part VI).

C4. Known Failure Modes

- Invented facts, quotations or sources, especially in science and social studies content.
- Misaligned standards: codes that look plausible but don't match the task.
- Arithmetic and answer-key errors in math.
- Lost rigor: "differentiated" versions that quietly lower the cognitive demand of the task.
- Drift from the curriculum: material that's reasonable in general but isn't tied to the adopted lesson.
- Generic or stereotyped examples, names and contexts.

Build Day practice, coaching and the quarterly output audit all focus on catching these.