

From Evidence to Instructional Practice

Four Lenses for Understanding Achievement Network's
Assessment Design

Eric Tucker (The Study Group) and
Michelle Odemwingie (Achievement Network)



From Evidence to Instructional Practice

Four Lenses for Understanding Achievement Network's Assessment Design

Eric Tucker (The Study Group) and Michelle Odemwingie (Achievement Network)

Abstract

How can assessment help a teacher decide what to teach next, and how can educators know whether that decision helped students learn? We examine Achievement Network's (ANet's) work through four lenses drawn from Robert J. Mislevy: practice, feedback, evidentiary argument, and measurement. These lenses connect the routines in which educators use evidence to inform the interpretations, decisions, and representations those routines support.

The experiences of Madison Metropolitan School District, Carlsbad Municipal Schools, and Honey Dew Elementary School show why useful assessment depends on curriculum coherence, professional learning, and teachers' knowledge of students. They also show where the chain can break: a report may arrive without time to interpret it, an interpretation may lose the context of student work, or a plausible instructional response may lack follow-up evidence.

We trace how these lessons inform ANet's current assessment design. Compass Math connects evidence of prior learning with anticipated needs in upcoming curriculum, bringing assessment closer to the teacher's next decision. Throughout, we distinguish evidence supporting an interpretation from evidence that an instructional action improved learning. The cases illustrate how schools organize this work; they do not establish an independent causal effect of an assessment.



When Evidence Loses the Learner as It Travels

David, a fifth grader in Honey Dew, offers a sentence with a design constraint inside it: "It's important for your teacher to know what you need to learn" (Odemwingie & Cockrell, 2025, p. 280). He is asking what the score is for. If his teachers do not learn what he needs, the test may have completed its measurement while missing its assignment.

That sentence changes the unit of analysis. A test can be studied as an object: standards, items, timing, score report, and technical properties. Its evidentiary meaning depends, however, on the use case in which those features are assembled. In his discussion of rich performance tasks, Mislevy asks whether adding depth, context, and interactivity improves assessment practice. The answer he emphasizes depends on the claim being made, the learner's experience, and what assessment users know about the instructional context. He therefore writes: "In certain assessment use cases, the answer is a resounding yes; in others, an emphatic no" (2017, p. 4). The question is often not whether a task is intrinsically meaningful, but what kind of learning claim the surrounding system can responsibly support.

We can therefore follow the chain that begins when a student answers and ends when a teacher, team, or student does something different. The first view asks whether an instrument measures well. The second asks whether an institution can learn.

Schools can have plenty of evidence and still miss the next instructional insight. Assessments may arrive too late, at the wrong grain size, or detached from the curriculum students encountered. Items may sort right from wrong without revealing reasoning. Reports may compress student work into a color or percentage, leaving teachers to supply meaning and the next move.

We examine ANet's work because it connects assessment with curriculum alignment, item study, reporting, coaching, professional learning communities (PLCs), walkthroughs, and student reflection. These connections matter when they help educators understand students' learning and decide how to respond. They also give us a way to examine what assessment design must provide, and what teachers and schools must contribute (Achievement Network, 2023b).

Our argument is that assessment becomes useful when a learning question, an interpretation, an instructional response, and follow-up evidence remain connected. ANet's experience helping schools make those connections is now informing the design of assessments intended to bring evidence closer to teachers' next decisions. We use four of Mislevy's frames—practice, feedback, argument, and measurement—to examine this work. The broader genre frame concerns the institutional purposes, expectations, and consequences that shape assessment use (Bell & Mislevy, 2021; Mislevy, 2012, 2018).

Madison, Carlsbad, and Honey Dew let us examine this work at different scales. Madison shows a district deciding which assessments deserve instructional time. Carlsbad shows how curriculum-aligned assessments, walkthroughs, and professional learning support consistent use of adopted materials. Honey Dew brings the question back to a student: what can the teacher learn from a response, and how can that knowledge guide the next lesson? We use the four lenses to examine the decisions each case makes visible and the evidence those decisions require (Mislevy, 2018, pp. 435–446; Odemwingie & Cockrell, 2025).

Four Lenses, One Assessment-Use Argument

The four lenses help us ask different questions about the same assessment work.



Practice directs attention to the people, relationships, routines, and resources through which an assessment is designed and used.



Feedback asks who needs information, for which decision, and when.



Evidentiary argument examines why a response supports an interpretation and what other explanations remain.



Measurement asks how scores and other representations can summarize patterns and uncertainty for a specified purpose.

(Bell & Mislevy, 2021; Mislevy, 2012, 2018, pp. 435–443).

These questions also guide design. The decision a teacher needs to make shapes the evidence an assessment must elicit and the form in which results should be reported. We therefore begin with the work educators and students need assessment to support.

The lenses are nested because each later question depends on meanings established earlier, and because intended use reaches backward into design. In their account of the duplex design, Bock and Mislevy write that “the different intended uses of the results influence the measurement procedures by which the data are collected and the statistical methods by which they are analyzed” (1986, p. 3). They begin not with a score but with the anticipated users, the decisions those users will make, and the information those decisions require.

We use four lenses drawn from Mislevy to examine ANet’s work in the same way. Practice identifies the work the assessment must enter; the feedback loop identifies who needs which representation; the evidentiary argument specifies what may be inferred; and measurement asks whether the resulting representation is scientifically sound for that use. A model cannot decide the instructional purpose; a report cannot supply the capacity to act; a response does not interpret itself.

We use “validity for use” as a chapter-level synthesis, not as a replacement for Mislevy’s definition of validity. For this chapter, an assessment-use chain is valid to the degree that evidence and theoretical rationales support the adequacy and appropriateness of its inferences and actions for a specified purpose. That chain runs from student response to interpretation, collective judgment, feasible and authorized instructional action, and follow-up evidence. The final link can strengthen or weaken the use argument, but a beneficial outcome alone does not establish that the original interpretation was warranted (Mislevy, 2018, pp. 201–209, 223–224).

Assessment as Practice: The Routines That Make Evidence Usable

The practice lens asks: What work must an assessment enter before its evidence can become usable? It moves attention away from the test event and toward the curriculum, routines, relationships, and decisions that give the event meaning. At ANet, that work begins before administration. Curriculum alignment, text selection, item writing, and item study ask whether a task represents the kind of thinking instruction is meant to develop.

This idea relates to what Behrens, Mislevy, DiCerbo, and Levy describe as the practical contribution of evidence-centered design (ECD): a common language for making visible the link that carries reasoning “from observation to conclusion,” even when that link is partly informal or intuitive (2010, p. 9). In that sense, alignment is not a one-time match between a standard and an item. It is a continuing effort to coordinate curriculum, task demands, scoring, reporting, and instructional use.

In English language arts, ANet uses a text-first approach and attends to the cognitive load of idioms and figurative language, a design concern with particular force for multilingual learners (Kodama, 2024). In mathematics, the distinction among conceptual understanding, procedural fluency, and application keeps a score from pretending that every form of competence is the same. These are design practices, not proof that every resulting inference is sound.

ANet’s open-book approach gives educators access to the reasoning behind tasks and answer choices. Mislevy and Riconscente argue that useful evidence must be represented in a form people can reason with and suited to its purpose (2005, p. 3). For a teacher, that may mean examining why an item was written, what an incorrect answer could suggest, and what to look for in a student’s work. A student reflecting on an answer, a PLC studying a class pattern, and a district allocating support need different representations of the evidence.

We treat misconceptions as clues to reasoning that teachers can investigate. A distractor rationale offers a possible explanation for an answer; it becomes instructionally useful when teachers compare that explanation with student work and connect it to the curriculum (Achievement Network, 2025d). This is the purpose of moving from a report of what happened to questions about why it happened and what to do next (Achievement Network, 2025c). Access to item rationales supports this inquiry, although it does not make every statistical model behind a report directly inspectable by educators.



That activity does not end at the report. Teachers, coaches, and leaders bring the evidence into walkthroughs, planning, and professional learning communities. A report is an object around which people can compare student work, challenge an interpretation, and decide what to try. Freeport School District's assistant superintendent, Georgia Mathis, calls PLCs "the lifeblood of schools" and describes them as teachers leading because teachers are closest to students and to the problem (Achievement Network, 2025b). Coaching adds another condition: professional judgment develops in conversation. One leader recalled learning with staff from a coach; another described a coach who left the work to the educator rather than taking it over (Achievement Network, n.d.).

Student reflection completes the practice without making students responsible for the system. If feedback is only a number, it remains an adult verdict. A Belle Chasse leader put the test plainly: "If we don't value feedback, and our kids don't value feedback beyond a number... then it's useless" (Achievement Network, 2025a). Student voice, representation, and language access therefore belong in the design question. Research on assessment tools points to student-facing utility and actionable reporting (Hoffmann et al., 2022); work on pedagogy and identity warns that how learning is taught, and whose identity is recognized, affects what participation can mean (Cipriano et al., 2025; Zieher et al., 2024).

The practice is consequential because it structures attention. It tells adults whether to look at a percentage, a piece of reasoning, or a pattern across students. It tells students whether an error is a verdict or an opening for the next lesson. But a routine can exist on paper and still fail in use. The next lens therefore asks: When evidence leaves one setting and enters another, who receives it, what do they need to know, and what decision can they responsibly make?

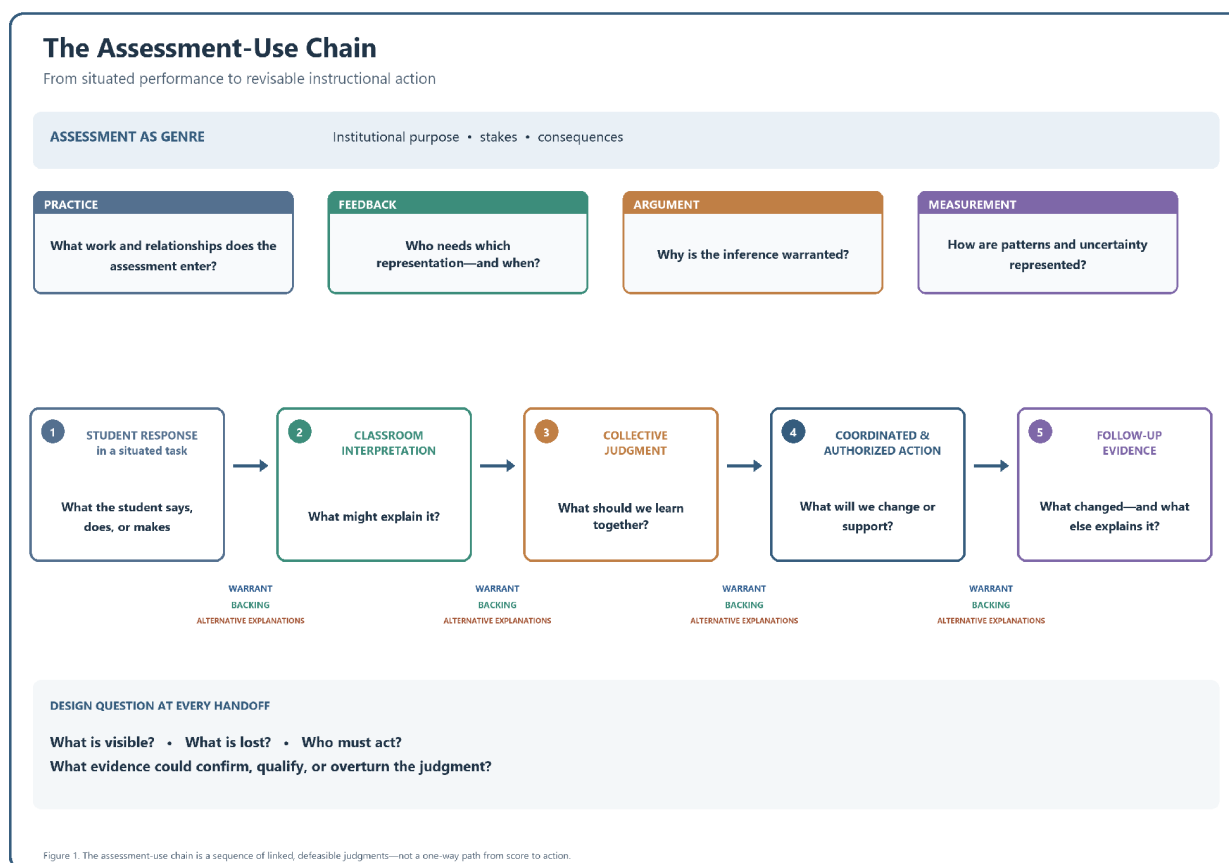
Feedback Loops: How Evidence Travels, Changes Meaning, and Closes

For this chapter, we treat feedback as a linked set of decision cycles. Each cycle has an actor, a question, a time scale, a grain of evidence, a plausible action, and a way to gather the next piece of information. The same task and performance can be useful for one purpose and weak for another because users know different things about the learner and situation. We treat a report as closing a feedback loop only when someone can connect the evidence to a plausible response and then revisit the learner's subsequent work (Mislevy, 2018, pp. 441–442; Oliveri et al., 2021, pp. 78–80).

Figure 1 makes the pathway visible: a student response enters a teacher's interpretation; that interpretation may enter a PLC's collective judgment; a school may aggregate the pattern into a leadership priority; a district may turn it into strategy, resources, or professional learning; and those decisions must return to classrooms as conditions for another round of work.

The same response can support different inferences for different users because the user's knowledge is part of the argument. Mislevy (2018) distinguishes the data in a student's performance from the additional information about the student's prior experiences and relation to the task; that information can strengthen a warrant or reopen an alternative explanation (pp. 70–79). This distinction clarifies why an ANet report must be interpreted inside the teacher's knowledge of a learner and curriculum, not treated as self-interpreting information.

Figure 1. The assessment-use chain from student response to instructional action



Note. The genre frame identifies the institutional purpose, stakes, and consequences that shape assessment use. The four lenses pose questions about the chain. Each transition requires reasons for the next judgment and attention to alternative explanations. Follow-up evidence can confirm, qualify, or overturn an earlier judgment.

At the classroom level, the useful report is not the one with the most fields. It is the one that helps a teacher move from what happened to why it may have happened and what to do next. Item-level information can support an instructional hypothesis, but only when joined to student work, knowledge of the task, curriculum history, and the teacher's understanding of the learner. The same item pattern can be a prompt for discussion in a PLC and an inadequate basis for a district mandate.

The team loop changes the work again. In a PLC, teachers can compare interpretations, ask whether a misconception is common or local, and decide whether to adjust a lesson, revisit a representation, or gather more evidence. Coaching can supply a second set of eyes without removing local ownership. The data-literacy literature names a similar progression as descriptive, diagnostic, and prescriptive insight: what happened, why it happened, and how a school might respond (Ng et al., 2025). Data literacy is therefore an organizational capability, not a trait that a dashboard can install.

Madison shows what happens when the loop reaches the district level. The district recognized a “glut of evaluative assessments” and moved toward measures closer to instruction (Achievement Network, 2021), consistent with ANet’s rule: “As few as possible. As formative as possible” (Achievement Network, 2023a). The audit created shared direction and a common question about which evidence deserved teachers’ time. Standardized reporting and PLC-based interpretation helped the district coordinate around instructional priorities.

Aggregation made something possible: a district could see redundancy, compare patterns, allocate support, and protect a common purpose. It also changed what was available to see. Item-level reasoning, classroom history, subgroup differences, and local explanations can disappear when only a school or district summary travels upward. Research on multilevel data makes the trade-off explicit: student-level information can identify needs, teacher-level information can guide support, while school-level information can hide group differences without sufficient breakdowns (Ng et al., 2025).

Aggregation changes what a report can support. A district summary may help leaders identify a shared instructional priority while offering too little detail to explain an individual student’s reasoning. At each move from classroom to team, school, or district, we therefore ask what the new user can decide, what detail has been lost, and what support must return to classrooms. Here, “grain size” includes both the level at which evidence is summarized and the detail needed for the intended decision.

A district decision becomes useful to teachers when it returns as time, curriculum support, coaching, professional learning, or a routine that helps them act. The earlier ANet case study describes a central lesson from program evaluation: access to data was insufficient when schools lacked the capacity to use it (Odemwingie & Cockrell, 2025, p. 278). This lesson connects the feedback lens to a design question: what can an assessment make easier, and what support must the school provide?



The distinction is visible in an early California assessment report by Mislevy and Bock, which says that statewide testing must provide results at a level of detail “sufficient to guide improvement” and must make it possible to examine the strengths, weaknesses, and trade-offs of alternative programs (1984, p. 2). Reporting detail is therefore not a matter of adding fields to a dashboard. It is a design judgment about what educators need to see in order to compare interpretations, choose a response, and allocate time and support. ANet’s reporting is strongest when its granularity is tied to that next decision. For evidence to remain answerable to learning, every upward movement must be paired with a credible path back down.

For this chapter, a feedback loop is complete when an assessment observation supports an interpretation, that interpretation supports a plausible and authorized instructional response, and subsequent student work supplies evidence that can confirm, qualify, or overturn the judgment. Mislevy’s distinction between assessment and criterion situations matters here: apparent similarity is only part of the warrant, because differences in practice and learner history can change what a response forecasts (Mislevy, 2018, pp. 105–111, 128–132).

Evidentiary Argument: From Response to Action to Consequence

The evidentiary-argument lens asks what must be true for a response to justify a decision. We distinguish the argument supporting an interpretation of student performance from the argument supporting its instructional use. The first connects a response to a claim about what a learner knows, can do, or may need next. The second asks why a particular instructional response is appropriate and what evidence would help educators judge its consequences (Mislevy, 2018, pp. 67–87, 105–132).

ANet’s open-book reports and collaborative routines make parts of this reasoning available for discussion. Teachers can examine item rationales, compare student work, question an interpretation, and decide what to investigate or teach. At each step, a warrant supplies the reason for moving from evidence to a claim or decision; alternative explanations show where that reason may be insufficient (Mislevy, 2012).

Four questions keep the links visible.

What does the performance show?

A response is situated. Mislevy and Durán argue that assessment should connect “assessment purposes, tasks, and cognitive and linguistic targets with authentic classroom experiences as curriculum practices aligned with standards” (2014, p. 561). For a multilingual learner, an incorrect response may reflect the interaction of language demands, topic familiarity, prior instruction, and the form of participation—not simply the absence of the targeted subject-area capability. The teacher’s knowledge of classroom work is therefore not background color; it is part of the information needed to interpret what the response can support.

We therefore consider three sources of evidence: what students say, do, or make; the features of the task; and students’ prior experiences in relation to that task. Language, representation, prior instruction, pacing, or attendance may help explain a response alongside the targeted knowledge or skill. A distractor rationale or diagnostic result should prompt teachers to test an interpretation against student work and classroom experience. Making this contextual knowledge explicit can strengthen an inference or reveal an alternative explanation that remains unresolved (Mislevy, 2018, pp. 73–79, 220–222; Mislevy, 2024, pp. 71–72).

This matters for equity because context is not an afterthought. Idiomatic language can raise cognitive load for a multilingual learner; a broad comprehension result can conceal decoding, fluency, confidence, or a student’s sense of being a reader. A common instrument may still require contextual knowledge to be fair and useful; without that knowledge, equity is compromised at the moment of interpretation. Identity and access affect how evidence is produced and understood.



What should educators infer?

State the inference as a bounded claim. “Several students chose the distractor associated with comparing denominators” describes an illustrative response pattern. “These students need a lesson using a number line” adds an instructional judgment. That judgment may be reasonable, but it needs support from the task, prior instruction, student work, and other evidence. We locate responsibility for the interpretation here: educators must be able to explain why the evidence supports their conclusion and what could lead them to revise it.

What instructional action should follow?

Interpretation and use are different arguments. A teacher may infer that students are struggling with a concept and still decide whether to reteach, model, practice, confer, or investigate. That decision depends on curriculum, time, pedagogy, relationships, and authority. Madison’s PLCs give teachers a place to decide together. Carlsbad paired formative assessments aligned to its adopted materials with instructional walkthroughs and targeted professional learning (Odemwingie & Cockrell, 2025, pp. 268–269). Coaching adds a social condition for judgment: leaders can acknowledge blind spots, try a small change, and retain ownership.

At the organizational level, the warrant must also authorize coordination. A pattern in one classroom may become a team problem when it recurs across student work and instruction. A team pattern may become a school priority when leaders can connect it to curriculum, time, and support. A district strategy is warranted only if its broader grain still serves an instructional question. Aggregation can justify shared direction; it cannot silently preserve the detail that aggregation removed. The organization must say what it knows, what it no longer knows, and what evidence would justify the next decision.



evidence supporting an interpretation



evidence that an instructional action
improved learning

What evidence indicates that the action mattered?

We distinguish three questions: Did the intended action occur? Did student performance or learning change? What evidence connects the action to that change? Evidence that a routine was implemented or that adults learned to use it supports the mechanism, but it does not close the causal argument. In Mislevy’s terms, evidence that the interpretation was warranted and evidence that the use produced beneficial consequences are related but different lines of reasoning. Honey Dew’s outcomes belong at the end of this chain: they are evidence to interpret, not proof that the assessment caused the change (Mislevy, 2018, pp. 201–209, 223–224).

Honey Dew's reported results include a 12.2 percent positive change in mathematics proficiency against a matched comparison and growth of 10 percent or more on 38 percent of assessed standards year over year (Odemwingie & Cockrell, 2025, pp. 279–281). These are different indicators of change. They warrant examination alongside the teaching-and-learning cycle, coaching, and other conditions at the school; they do not, by themselves, isolate the effect of the assessment.

The stronger account asks what happened between the response and the result. Honey Dew's teaching-and-learning cycle had to be learned and revised. One teacher moved from "I straight up did not understand the point" to describing the process as beneficial to her and her students (Achievement Network, 2024). That is evidence of adult learning and uptake. It supports a mechanism, but it does not close the causal argument. Curriculum coherence, leadership, coaching dosage, teacher learning, student mobility, timing, selection, and other concurrent changes remain plausible alternatives.

Validity for use therefore names a chain of warrants, not a property that belongs to a score alone. The response must be interpretable for a legitimate learning purpose; the inference must support a feasible and authorized action; and the action must produce further evidence that can confirm, qualify, or overturn the original judgment. The measurement lens asks how tools help represent that reasoning without pretending to replace it.

Measurement: Provisional Models for Instructional Visibility

Measurement is our final lens because the meaning of a score depends on the work it is meant to support. Scores, growth estimates, and diagnostics represent patterns in limited observations. Their value depends on whether the evidence and assumptions are adequate for the intended inference, and whether users recognize uncertainty and possible misfit (Mislevy, 2018, pp. 156–159, 177–179; Mislevy, 2024, pp. 71–72).

Mislevy distinguishes the target of an inference from the evidence gathered for it and the methods used to judge that evidence (2003). This distinction helps us separate two kinds of scrutiny. Educators can inspect student work, item rationales, and the instructional meaning of a report. Compass Math's predictive model requires technical evaluation of its performance. Its accuracy and calibration are evaluated against later responses; learning components with thin evidence are reported only in aggregate and flagged as having limited data. Teachers can then consider the reported pattern alongside what they know about their students, without being expected to inspect the model itself.

Precision is not the same as usefulness. Mislevy distinguishes broader reliability from familiar internal-consistency indices, then examines the relevance, credibility, and inferential value of evidence for an intended inference (2004, pp. 241–244). A report can be numerically precise and still be poorly suited to the question a teacher is trying to answer; conversely, a set of non-interchangeable observations may be valuable when the inference concerns a particular performance or learning episode. The measurement question is therefore not simply how stable a number looks, but whether the evidence is relevant, credible, and strong enough for the decision that follows. A representation that helps a teacher ask a local question may be too coarse for district comparison; a district summary may be too blunt for diagnosis. The test is fit with purpose.

These distinctions matter for current assessment design. Compass Math connects assessment evidence with the curriculum teachers are preparing to teach. Evidence about the accuracy and calibration of its predictions addresses one question; evidence about effects on instruction and student learning addresses another. We need to state which claim each source supports. Kiddom’s Atlas, developed with ANet and Teaching Lab, embeds assessment within instruction; ANet contributed the curriculum-aligned assessment engine (Achievement Network, 2026). Both examples bring us back to the teacher’s decision: what does the evidence help the teacher see, and how does it support a response within the adopted curriculum?

Stanford’s Rapid Online Assessment of Reading (ROAR) illustrates why the target of a measure matters. ANet piloted ROAR with Reading Reimagined to examine foundational reading skills that a broad comprehension result might leave unclear (Odemwingie & Cockrell, 2025, pp. 275–276). In that pilot, a separate Motivation to Read Profile addressed reading self-efficacy. Together, the measures let educators consider reading skills and students’ experience of reading without treating one score as an explanation of both.

The same questions apply to adaptive diagnostics, growth estimates, and item-level patterns:



What is the target?



What observations feed the representation?



Which assumptions connect response to interpretation?



Where might context or identity-linked differences be lost?



What action does the tool invite, and what evidence will test that action?

Practical validity requires the report to become meaningful to students and educators, not merely technically defensible (Ng et al., 2025).

Technology cannot answer those questions by itself. Mislevy, Almond, Yan, and Steinberg name the underlying problem plainly: educational assessment draws inferences about what a student knows or can do from limited observations of what the student says or does (1999, p. 437). Bayesian models can update beliefs as new evidence arrives, but the model does not decide what the evidence means, which observations matter, or what action a school should take. The authors’ closing lesson is organizational as much as statistical: complex assessment requires psychology, technology, statistics, and the substantive domain to be coordinated from the beginning (p. 446).

We see technology as useful when it helps teachers find relevant patterns and prepare an instructional response. The design goal is to strengthen teaching by reducing the work between evidence and a well-informed decision (Golston, 2025). Teachers still need to interpret student work, consider the curriculum, and examine what happens after they act. A report can support that process, but its contribution must be judged within the teaching and learning it helps make possible.

Madison, Carlsbad, and Honey Dew: One Assessment Ecology

The three cases show where decisions about evidence meet the conditions for acting on it. Madison concerns assessment selection and shared instructional purpose. Carlsbad concerns fidelity to adopted materials and the professional learning needed to support it. Honey Dew concerns teachers learning to use evidence within a teaching-and-learning cycle. Across the cases, we ask what the evidence made visible, what alternative explanations remained, and what educators had the support and authority to do next.

Madison began by examining which assessments served a shared instructional purpose. District leaders described a “glut of evaluative assessments”; the audit, reporting work, and PLC routines made assessment selection a district responsibility (Achievement Network, 2021). ANet’s rule, “As few as possible. As formative as possible,” is useful here because it asks every measure to earn its place in the instructional calendar (Achievement Network, 2023a). The case reports substantial changes in school-leader perceptions of instructional priorities and data use. We read those findings as evidence of greater shared direction, while recognizing that perceptions and student outcomes answer different questions. Local educators remained responsible for interpreting evidence and deciding how to respond.

Carlsbad shows why adopting high-quality instructional materials does not ensure consistent implementation. Teachers were using supplemental resources, and pacing and rigor varied across classrooms. The district used curriculum-aligned formative assessments and instructional walkthroughs to identify gaps in implementation, directed professional learning toward those gaps, and created an instructional leadership department focused on coherence (Odemwingie & Cockrell, 2025, pp. 268–269). A principal’s observation that classrooms had achieved “coherence and consistency” describes a change in instructional conditions. It does not, on its own, establish a change in student learning.

Carlsbad’s version of coherence emphasized fidelity to adopted materials (Odemwingie & Cockrell, 2025, pp. 268–269). Compass Math carries that emphasis into assessment design: its reports direct teachers to specific lessons in the adopted curriculum rather than to supplemental materials. This connection gives teachers a curricular starting point for responding to evidence. Their professional judgment remains necessary in deciding how to use those lessons with the students in front of them.



Honey Dew brings the argument back to teachers learning to use evidence. One teacher's account of coming to value the process illustrates a change in understanding and participation (Achievement Network, 2024). We treat that account as evidence about how the teaching-and-learning cycle was taken up, alongside the school's reported results. It helps explain the work involved without isolating the assessment's contribution to those results.

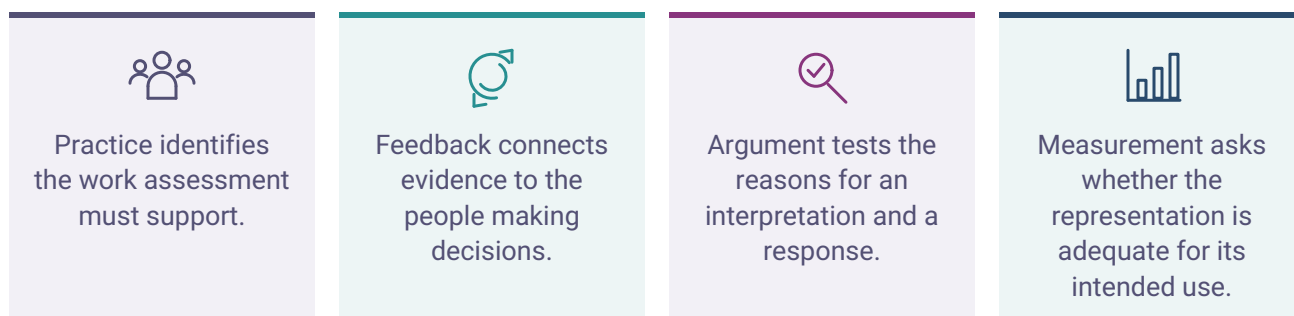
Together, the cases show why responsibility has to be shared across a school system. District leaders set priorities and provide time and support; school leaders and coaches help teachers use evidence; teachers bring knowledge of students and classroom work to the decision. We locate equity in these decisions: whether students can access the task, whether reports preserve meaningful differences, and whether teachers and students can use the evidence to guide further learning (UnboundEd, 2026). These are responsibilities for design and practice, not outcomes that can be attributed to a platform alone.

These experiences also sharpen the question for assessment designers: how much of the work between a student response and a teacher's next decision can the assessment itself support? The answer must draw on what schools have learned about curriculum, interpretation, and the time teachers need to act.

When Evidence Remains Answerable to Learning

Return to David's sentence and the standard it sets. A teacher needs to know what a student needs to learn. For that knowledge to matter, the system must preserve more than a score as evidence travels. It must preserve a learning question, a defensible interpretation, a possible action, and a way to see what happened next.

The four lenses help us examine how that happens.



Taken together, they help us examine the work schools do with assessment and the design choices that can make that work more useful.

The cases also identify work that assessment design cannot complete by itself. Educators need time to interpret evidence, knowledge of the curriculum and their students, and support for acting on what they learn. We must also ask whose experience is missing from a report and how much context disappears when evidence is summarized.

Two questions guide our judgment: Was the interpretation warranted, and what evidence shows whether the action helped? The first concerns the adequacy of the inference; the second concerns what followed the decision. Mislevy's account of validity requires attention to the adequacy and appropriateness of inferences and actions for a specified purpose (2018, pp. 194–209). For us, that means examining whether an interpretation fits student work, whether the proposed response is instructionally plausible, and whether subsequent changes have other reasonable explanations.

This accumulated experience is also shaping ANet's next generation of assessment design. After two decades of helping educators translate assessment evidence into instructional action, ANet is increasingly asking how assessment itself can shorten the distance between evidence and a teacher's next decision. Compass Math reflects that evolution. Its short, curriculum-compatible milestones look both backward and forward: surfacing misconceptions students may carry from prior learning while anticipating where those misconceptions may create difficulty in the curriculum ahead. By connecting those insights to the content teachers are preparing to teach, Compass is designed not only to explain what has already happened, but to help teachers prepare for what students are likely to need next.

We return, then, to David's expectation that his teacher should know what he needs to learn. Reports, coaching, item study, and PLCs help educators interpret evidence and use it. Assessment design can make that work easier by connecting what students show now with the lessons ahead. The standard we bring to both is whether the resulting decisions are well supported and help teachers respond to students' learning.



The test should serve instruction; instruction should not be made to serve the test.

References

- Achievement Network. (2021). *Creating district coherence through assessment strategy*. <https://www.achievementnetwork.org/resource-center/case-study-creating-district-coherence-through-assessment-strategy>
- Achievement Network. (2023a). *Assessing assessments*. <https://www.achievementnetwork.org/resource-center/assessing-assessments>
- Achievement Network. (2023b). *Our values for interim, formative assessments*. <https://www.achievementnetwork.org/resource-center/formative-assessment-values>
- Achievement Network. (2024). *Teaching and learning cycle and educational leadership: Honey Dew Elementary School case study*. <https://www.achievementnetwork.org/resource-center/teaching-learning-cycle>
- Achievement Network. (2025a). *Belle Chasse Academy: HQIM as a foundation*. <https://www.achievementnetwork.org/resource-center/belle-chasse-hqim>
- Achievement Network. (2025b). *Freeport school district—a model of HQIM implementation*. <https://www.achievementnetwork.org/resource-center/freeport-hqim>
- Achievement Network. (2025c). *Instructional utility: Using assessment to fuel instruction*. <https://www.achievementnetwork.org/resource-center/instructional-utility>
- Achievement Network. (2025d). *The power of misconceptions in instruction*. <https://www.achievementnetwork.org/resource-center/misconceptions>
- Achievement Network. (2026). *Atlas: What it takes to make assessment the engine of instruction*. <https://www.achievementnetwork.org/news/project-atlas>
- Achievement Network. (n.d.). *ANet partners on coaching* [Video]. YouTube. <https://youtu.be/oEOQuHnVj2w>
- Bell, C. A., & Mislevy, R. J. (2021). Practice, feedback, argument, measurement: A frame for understanding diverse perspectives on teaching assessments. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of analyzing teaching quality: Potentials and pitfalls* (pp. 21–52). Scandinavian University Press. <https://doi.org/10.18261/9788215045054-2021-01>
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world* (CRESST Report 778). National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
- Bock, R. D., & Mislevy, R. J. (1986). *Comprehensive educational assessment for the states: The duplex design* (CSE Report No. 262). Center for the Study of Evaluation, University of California, Los Angeles.

- Cipriano, C., Ahmad, E., McCarthy, M. F., Ha, C., & Ross, K. (2025). Illustrating the need for centering student identity in universal school-based social and emotional learning. *Social and Emotional Learning: Research, Practice, and Policy*, 5, 100088.
- Golston, A. (2025, October 7). *What it looks like to design technology for teachers and the difference it makes in the classroom (featuring Michelle Odemwingie)*. Gates Foundation U.S. Program Insights. <https://usprogram.gatesfoundation.org/news-and-insights/articles/what-it-looks-like-to-design-technology-for-teachers-and-the-difference-it-makes-in-the-classroom>
- Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment*. Educational Testing Service.
- Hoffmann, J. D., Foster, B., Hills, K. J., Bradley, C., Cipriano, C., & Brackett, M. A. (2022). Building useful, web-based educational assessment tools for students, with students: A demonstration with the school climate walkthrough. *Assessment in Education: Principles, Policy & Practice*. <https://doi.org/10.1080/0969594X.2022.2056576>
- Kodama, M. (2024, April 29). *Japanese and English bilingualism in ELA assessment development*. In Questions about multilingual learners with Dr. Ayanna Cooper: Part 5. Achievement Network. <https://www.achievementnetwork.org/resource-center/ml-el-blog-5>
- Mislevy, R. J. (2003). Rehabilitating psychometrics: Commentary on Pellegrino and Chudowsky's "The foundations of assessment." *Measurement: Interdisciplinary Research and Perspectives*, 1, 162–165.
- Mislevy, R. J. (2004). Can there be reliability without "reliability?" *Journal of Educational and Behavioral Statistics*, 29(2), 241–244. <https://doi.org/10.3102/10769986029002241>
- Mislevy, R. J. (2012). *Four metaphors we need to understand assessment*. Gordon Commission on the Future of Assessment in K–12 Education. https://www.gordoncommission.org/rsc/pdfs/mislevy_four_metaphors_understand_assessment.pdf
- Mislevy, R. J. (2017). Resolving the paradox of rich performance tasks. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 1–46). Information Age Publishing.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J. (2024). Sociocognitive and argumentation perspectives on psychometric modeling in educational assessment. *Psychometrika*, 89(1), 64–83. <https://doi.org/10.1007/s11336-024-09966-5>
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design (Research Report No. RR-03-16). Educational Testing Service.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

- Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In K. B. Laskey & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). Morgan Kaufmann Publishers.
- Mislevy, R. J., & Bock, R. D. (1984). *A technical description of the procedures used in calculating school-level scaled scores for the "Survey of Basic Skills: Grade 6."* California State Department of Education.
- Mislevy, R. J., & Durán, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of Common Core and Next Generation Science Standards. *TESOL Quarterly*, 48(3), 560–585.
<https://doi.org/10.1002/tesq.177>
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
<https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). SRI International.
- Ng, Z. J., Foster, B., Ellis, T. L., Barnes, S. P., Hoffmann, J. D., & Cipriano, C. (2025). Promoting data literacy: Using social and emotional learning assessment data to inform teaching and learning. *Behavioral Sciences*, 15(12), 1728. <https://doi.org/10.3390/bs15121728>
- Odemwingie, M., & Cockrell, K. (2025). From evaluation to impact: Transforming assessment into a tool for learning. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning* (pp. 245–282). University of Massachusetts Amherst Libraries.
https://www.assessment-for-learning.org/_files/ugd/e901ef_12994eef6f484597bd15565bce8940ff.pdf
- Oliveri, M. E., Slomp, D., Rupp, A. A., & Mislevy, R. J. (2021). Principled development of Workplace English Communication Part 2: Expanded evidence-centered design and theory of action frameworks. *Journal of Writing Analytics*, 5, 71–108. <https://doi.org/10.37514/JWA-J.2021.5.1.03>
- Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Caliço, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, 4(1), 49–102.
<https://doi.org/10.5281/zenodo.3554643>
- UnboundEd. (2026). *Michelle Odemwingie*. <https://unbounded.org/people/michelle-odemwingie>
- Zieher, A. K., et al. (2024). Considering the “how” of SEL: A framework for the pedagogies of social and emotional learning. *Social and Emotional Learning: Research, Practice, and Policy*, 3, 100030.