

The Public Life of a State Test

New Meridian and Montana's MAST: Using Evidence Centered Design to Improve Public Debate Regarding Consequential K-12 Assessment

Eric Tucker



The Public Life of a State Test

New Meridian and Montana's MAST: Using Evidence Centered Design to Improve Public Debate Regarding Consequential K-12 Assessment

Eric Tucker

Abstract

Public arguments about K-12 assessment often compress four different questions into a single debate:



Can the system
produce a defensible
score?



Can people benefit
from its reports?



Does it fit the
instructional settings
in which students
learn?



Is the burden worth
the public value?

Because these questions rely on different evidence and different standards of proof, stakeholders can talk past one another while appearing to debate the same test.

This case study examines New Meridian's work with the Montana Office of Public Instruction on the Montana Aligned to Standards Through-Year (MAST) assessment for grades 3–8 English language arts and mathematics. Its first operational year produced testlet reports during the year and a statewide summative result. Those products support different decisions, and their justification requires different evidence.

Drawing on Robert Mislevy's roots in Toulminian argumentation, the case study proposes Evidence-Centered Design as a paradigm for judging public debate about K-12 assessment systems. Framed through Mislevy's account of assessment as practice, feedback, evidentiary argument, and measurement, the chapter reconstructs some of the reasoning a consequential assessment system must make public. It distinguishes design and interpretation arguments from use arguments; separates data from evidence; surfaces warrants, qualifications, and alternative explanations; and identifies studies needed to move from promising architecture to defensible use. Its four-part test is straightforward to state but demanding to meet: prove the score, the report, the setting, and the balance between burden and public value.

The chapter's contribution is a way to make public disagreement precise and consequential. Participants should identify the inference under dispute, acknowledge the evidence that supports it, and specify what would strengthen, narrow, or overturn their judgment. Validation continues as designs and uses change. That discipline also matters when artificial intelligence (AI) adds new forms of scoring, feedback, or recommendation, each with its own claims to justify.

The Sentence That Leaves Usefulness Open

In its first operational year, 2024–2025, MAST put a through-year design into statewide use in grades 3–8 English language arts and mathematics. Students encountered three operational windows, while a fourth supported linking and field testing. The system produced testlet reports during the year and an aggregated summative score. Its technical report documents the work behind those products, including linking, reliability and classification analyses, scoring controls, and fairness studies (Montana Office of Public Instruction & New Meridian, 2025).

That record gives the score argument something substantial to examine. It also identifies a further line of inquiry: “In future years, New Meridian will collect and analyze more data regarding the usefulness of the testlet reports” (Montana Office of Public Instruction & New Meridian, 2025, p. 183). The sentence directs attention to a different claim: what educators can do with the information the system provides.

The distinction is between two validity arguments. One concerns the defensibility of the score for a specified interpretation and use. The other concerns whether a report helps an educator make an appropriate decision while there is time to act. An accurate report may still be confusing, too broad, or disconnected from a feasible response. A report that feels useful may invite an interpretation its evidence cannot support.

MAST makes these questions concrete because it has schedules, reports, psychometric choices, and people who must work with them. Its technical record and implementation experience allow each claim to be examined and revised. The task is to judge the support for particular uses now and identify what further evidence would change that judgment.

Why MAST matters

MAST addresses a familiar assessment problem: states need comparable achievement information, while educators need information soon enough and at a useful level of detail to inform instruction. Huff (2025) places that tension within a broader demand for assessments that address complex constructs and multiple purposes without simply multiplying tests. MAST offers a concrete design in which to examine how those purposes interact.

Through-year systems differ in what they assess, how observations contribute to summative results, and how they relate to local curriculum (Ismail et al., 2023; Levitan & Munyan-Penney, 2024; Olson, 2024). This case therefore examines MAST’s particular claims. It does not independently re-estimate its psychometric models or establish causal effects. It asks what the public record supports, what conditions qualify that support, and what studies would strengthen or challenge it.



One Assessment System, Several Publics

MAST's users ask different questions. State leaders need defensible achievement interpretations. Teachers need information that adds to classroom evidence and supports a feasible response. District leaders must reconcile local pacing with state expectations. Students and families experience the testing, the reports, and the consequences of decisions made from them. Mislevy's feedback-loop question keeps those differences visible:

“Who needs what information, when, for what reasons, and what other information do they have to work with?”

(2013, Section 2.2).



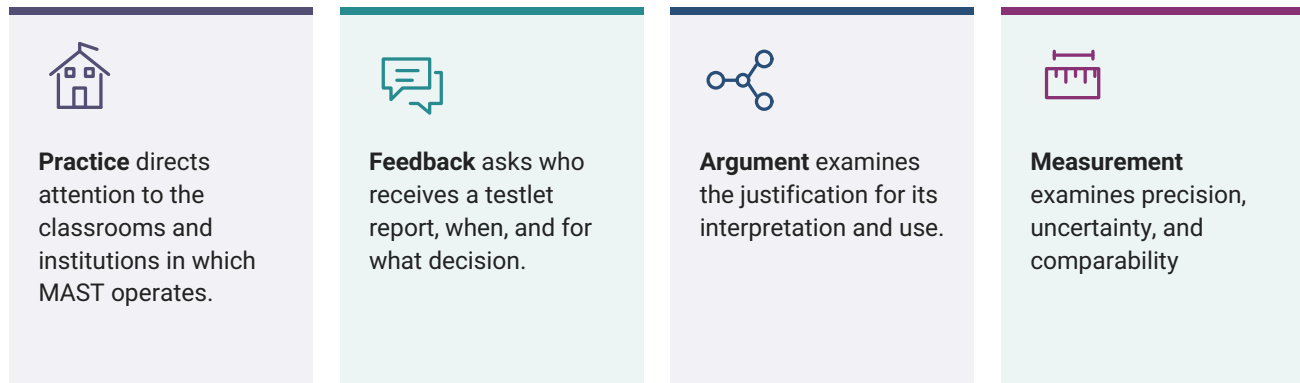
A public exchange makes the disagreement concrete. Chu (2026) argues that demands for comparability and instructional responsiveness create structural tensions that through-year designs may not resolve. VanderVeen (2026) acknowledges the risk of confusing the two purposes but argues that better design can serve both. He presents MAST as an effort to improve the coherence of instructional decision making and specifies that its testlets are not intended to replace classroom formative assessment or curriculum-embedded assessments.

The dispute turns on which design choices can reconcile the purposes, under what conditions, and with what evidence. Mislevy's lenses help separate those questions so that technical findings, classroom experience, and public judgments can each do the work for which they are suited (Bell & Mislevy, 2022).

An argument framework for the MAST case

Evidence-Centered Design (ECD) begins with the claims an assessment should support, the evidence those claims require, and tasks that can elicit that evidence (Mislevy et al., 2003). Applied to MAST, it asks how performances across testlets support an achievement interpretation and how that interpretation supports a particular decision. The second argument introduces questions about users, available actions, and consequences that score evidence alone cannot answer (Mislevy, 2018).

Mislevy's four lenses keep the case in view.



(Bell & Mislevy, 2022; Mislevy, 2013).

Toulmin's structure makes that reasoning explicit. The claim states what is proposed; data supply observations; the warrant explains why those observations support the claim. Research and theory provide backing, while qualifications and alternative explanations identify limits (Schum, 1994; Toulmin, 1958). MAST's reliability estimates, for example, bear on the precision of its summative scores. Their relevance to that claim does not establish that teachers understand testlet reports or make better instructional decisions because of them. Report-comprehension research bears on those later links. Evidence becomes useful when its relationship to a particular claim is clear (Mislevy, 2018).

The four arguments below therefore use the same structure: claim, current evidence, warrant and backing, qualifications and alternatives, and additional evidence needed. "Current evidence" means evidence discussed in this case, not an exhaustive inventory of studies. An unresolved question is a research need; an observed problem is evidence that may require a change in design or use. Both belong in an ongoing validity argument.

The analogy to a public-debate judging paradigm is limited but useful: participants should make their standards of relevance and comparison visible (Dudczak et al., 1992; Freeley, 1981; Lichtman & Rohrer, 1982; Rowland, 1982). ECD helps expose the reasoning; public judgment remains necessary to decide acceptable burden and consequences. The same discipline will be needed if artificial intelligence (AI) adds scoring, feedback, or recommendations to an assessment system, because each addition introduces claims that require their own justification.

Two Use Arguments Inside One Assessment System

MAST places two consequential use arguments inside one architecture. The first is familiar to state assessment. Students respond to tasks across several administrations; their performances are evaluated, linked, and synthesized into claims about achievement. Those claims may inform accountability judgments, classifications, interventions, or resource decisions. The technical questions (content representation, linking, reliability, classification accuracy, standard setting, security, accessibility, fairness, and stability across windows and groups) matter because the inferences will be leveraged for decisions affecting learners, schools, and systems.

The second use argument concerns assessment in the service of learning. Here the desired result is not merely a valid description of achievement. An early report must reach an educator before the instructional moment has passed, communicate its meaning accurately, add something to classroom evidence, and connect to a response the educator can take. The action must improve the next decision or student experience enough to justify the testing, reporting, training, and attention the system requires.



The arguments are distinct but not independent. The instructional report inherits assumptions about what tasks elicited and what a score or category means. The high-stakes interpretation can be affected by curricular sequence, opportunity to learn, and the behaviors the reporting system creates. Montana's March 2026 guidance tells educators to interpret progress reports alongside classroom evidence and not as final scores or replacements for teacher judgment (Montana Office of Public Instruction, 2026). That guidance sets a limit on how this particular report should inform decisions. If an early indicator becomes a stand-alone label, the use has changed, and the argument required to justify it changes as well.

The wager is consequently more demanding than the claim that testing more often is better. It is that one evidence system can sustain a defensible state interpretation and a useful instructional signal without allowing either use to borrow validity it has not earned from the other. The public record should be read against four arguments: score, report, setting, and burden relative to public value.



Argument One: The Score

Claim. Performances collected across MAST administrations can support fair and sufficiently precise achievement interpretations for specified state uses.

Current evidence. The first operational technical report documents item development, linking, model-based scoring, reliability and classification analyses, human-scoring controls, differential item functioning, standard setting, and security procedures (Montana Office of Public Instruction & New Meridian, 2025). These analyses bear directly on the score argument. Their relevance is stronger and more specific than the fact that the system produced scores.

Warrant and backing. The inference depends on the tasks representing the intended achievement domain and on the scoring and linking models accounting adequately for relevant differences across administrations. Content specifications, model-fit and linking analyses, scoring studies, and fairness analyses supply backing to examine that warrant. The question is how well their findings support the particular interpretation and population at issue, including the uncertainty that remains.

Qualifications and alternatives. Support for one score interpretation does not automatically extend to another. Precision sufficient for aggregate monitoring may be insufficient for an individual decision. Differences in opportunity to learn, curricular sequence, task familiarity, or response demands may affect what performance means across students and settings (Mislevy, 2018). These are questions to investigate against the technical findings, not findings that MAST's scores are necessarily incomparable.

Additional evidence needed. The next step is to connect the reported results to each intended state use, identify where existing analyses address variation across windows, sequences, districts, curricula, and student groups, and extend them where material uncertainty remains. Studies relating opportunity to learn to score interpretation would test a central alternative explanation. Public documentation should specify the supported uses, their qualifications, and uses the evidence does not support.



Argument Two: The Report

Claim. MAST reports, interpreted alongside classroom evidence and appropriate professional support, improve educators' instructional decisions. Claims about subsequent learning require evidence that follows those decisions into student experience and outcomes. As VanderVeen (2026) puts it, "It is not enough to display data faster or more often" (para. 6).

Current evidence. Park's (2024) interviews using early MAST report mockups identified both promising features and interpretation problems. Participants saw potential instructional value in misconception information and links to standards and responses. Some misunderstood scale scores or found ELA clusters too broad; Park also noted that scale scores would no longer appear in testlet reports. Participants encountered mockups without test or administration context. These findings support specific design revisions and hypotheses about usefulness, while leaving operational interpretation and classroom effects to further study.

Montana's FAQ states that testlet reports are available weekly during testing windows (Montana Office of Public Instruction, n.d.-b). Availability supports the intended feedback cycle, although teachers' access and response still require examination. Education First's report on its implementation partnership with the Montana Office of Public Instruction finds that educators valued MAST's mathematics misconception information, while also identifying uneven use and problems with some instructional responses (Badrinarayan et al., 2026). VanderVeen (2026) reports favorable student and teacher survey responses. These sources contribute evidence about perceived value and implementation; they do not by themselves establish that MAST caused learning gains.

Warrant and backing. Useful feedback must arrive in time, convey a distinction the educator understands, add to available classroom evidence, and support an appropriate action. Report research and implementation findings help test those links. The use argument requires evidence beyond support for the achievement interpretation itself (Mislevy, 2018, p. 206).

Qualifications and alternatives. Reports may be too broad to guide instruction, add little to teachers' existing evidence, or encourage premature labels and disconnected remediation. Professional learning may be part of the mechanism through which reports help. It is a rival explanation only when a claim attributes the benefit to the report alone. Studies should therefore specify whether they evaluate the report, the support package, or the combined system, and consider how curriculum and other instructional changes contribute to the observed result.

Additional evidence needed. Follow operational reports from access and interpretation to the decisions they inform. Compare the resulting decisions with those educators would make using existing classroom evidence and supports. For claims about improved learning, examine student experiences and outcomes with designs that address plausible competing explanations. Persistence and transfer require delayed measures and tasks that extend beyond recently taught material.

The evidence can support progressively stronger claims: comprehensible reports, useful contributions to decisions, and improvements in learning. A finding at one stage should inform the next inquiry without being made to stand for it.



Argument Three: The Setting

Claim. MAST’s accommodation of local instructional sequence can coexist with fair and comparable state achievement interpretations. Its early reports should also be interpretable in relation to what students have had an opportunity to learn.

Current evidence. Montana’s public description identifies flexibility to align mathematics testing with local scope and sequence as a design goal (Montana Office of Public Instruction, n.d.-a). That goal makes curriculum and timing part of the case to examine, rather than incidental features of implementation.

An early MAST pilot identified mismatches between mathematics testlet order and what students had been taught (Ismail et al., 2023, p. 94). In his 2023 interview with Chu, VanderVeen described flexible sequencing as a design intended to accommodate local curricula while preserving comparable summative reporting (Chu, 2023). These sources document an implementation problem and a design intention; they do not by themselves show that sequencing flexibility resolved the mismatch. Its effects on instructional fit and comparability remain empirical questions.

Warrant and backing. Differences in sequence and timing must preserve the intended achievement claim or be represented adequately in its interpretation. Curriculum analysis and evidence about timing, exposure, and performance are needed to support that warrant. Badrinarayan’s (2024, p. 16) published framework distinguishes several ways to connect assessment with curriculum, including alignment to local scope and sequence and to specific instructional materials. It places Montana’s mathematics model in the former category.

Qualifications and alternatives. Familiarity with taught content may affect performance, early results may reflect exposure as well as capability, and testing windows may influence pacing despite formal local flexibility. Conversely, a uniform schedule may weaken instructional relevance. Each possibility calls for evidence about the particular MAST configuration in which it could arise.

The implementation record supplies concrete concerns. Badrinarayan et al. (2026, p. 12) describe Montana educators’ reports of adjusting pacing to testing windows and responding to isolated standards in ways that disrupted instructional coherence. Those accounts challenge the assumption that local flexibility alone secures instructional fit. They establish experiences that require investigation and response; they do not establish how widespread those experiences are or isolate the effects of MAST from other pressures.

Additional evidence needed. Relate testlet administration dates to curriculum and pacing records, then examine whether score interpretations and report usefulness vary with instructional sequence, district context, or student group. Study how educators use the available flexibility and whether it resolves the earlier mismatches. Compare relevant curriculum relationships where alternative designs are feasible, including their costs and implications for transfer. The findings should identify which variations are compatible with a common claim and which require qualification or redesign.

The setting is part of the warrant. Professional learning, reporting language, and guidance on early results help determine whether the intended use is feasible in a particular school.



Argument Four: Burden and Public Value

Claim. MAST's contribution to educational decisions and student experience justifies its demands on time, resources, attention, and public authority relative to feasible alternatives. This is partly an empirical claim about effects and costs, and partly a public judgment about acceptable tradeoffs. Mislevy discusses evidence of student improvement alongside broader questions of cost, fairness, and washback (2018, p. 209).

Current evidence. MAST's administration and reporting arrangements identify recurring demands to study, and its implementation record describes experiences that may contribute to benefit or burden. Those sources do not supply a complete comparison of total costs and public value. Olson (2024) questions whether through-year testing reduces total testing time, while Levitan and Munyan-Penney (2024) emphasize equitable access to useful information and support for interpreting it. Their analyses help frame the comparison; they do not establish MAST's net burden.

Warrant and backing. An assessment system creates public value when its information supports worthwhile decisions and experiences at an acceptable cost. That judgment requires evidence about who benefits and who bears the work or risk. Testing time matters to students; administration, interpretation, and response matter to educators; technology, training, and governance matter to districts and the state. Differences in local capacity may change the balance. The comparison should include assessments and activities that MAST actually replaces as well as those that remain.

Qualifications and alternatives. The system may redistribute testing burden without reducing it, add reporting demands without capacity to respond, or produce benefits in some settings at disproportionate cost in others. Regulation can constrain innovation while protecting comparability, fairness, security, and technical quality (Ismail et al., 2023). Regulatory status should be reported for the specific use at issue; it does not by itself establish instructional usefulness or net public value.

Governance affects the credibility of this argument. New Meridian develops the assessment with Montana's Office of Public Instruction, while the state is responsible for public decisions about its use. Their evidence should be open to independent examination, with the roles of developer, state, and reviewer made clear. A provider's account is relevant evidence, and its claims remain subject to the same scrutiny as a critic's. That standard protects the public without presuming an issue or treating institutional affiliation as a verdict.

Additional evidence needed. Compare MAST with the prior assessment arrangements and feasible alternatives, accounting for tests or activities retained, replaced, or added. Measure total student testing time, staff work, costs, decision quality, and student and educator experience, including how these vary across schools and groups. Distinguish transition costs from recurring demands. The public decision should state which benefits and costs are established, which remain uncertain, and why the resulting tradeoff is acceptable or requires revision.

What Stronger Public Disagreement Would Require

The four arguments give participants a common task: name the claim at issue, acknowledge the evidence bearing on it, and identify what would change the judgment. A dispute about MAST's summative score calls for different evidence from a dispute about whether an early report improves a teacher's next decision. Both matter, and neither is resolved by a general verdict about through-year assessment.

Proponents should state the qualifications attached to their claims and publish findings that reveal problems as well as strengths. Critics should distinguish adverse findings from unanswered questions and identify evidence that could reduce their concern. States should specify permitted uses and supports, preserve qualifications in reporting and training, and authorize independent scrutiny. Researchers should connect the next study to an uncertainty that matters for a real decision.

Students, families, and educators contribute evidence about conditions and consequences that technical analyses alone cannot supply. Their experiences also inform judgments that evidence cannot settle by itself: how much burden is acceptable, how authority should be shared, and which consequences warrant public action. Mislevy's lenses help keep those responsibilities connected without making one kind of knowledge substitute for all the others (Mislevy, 2013).

Extending the argument to AI

The distinction introduced earlier becomes consequential if and when AI is proposed for scoring, feedback, or instructional recommendations. Agreement between automated and human scores bears on a scoring claim. It does not establish that a recommendation is instructionally sound, feasible in a particular classroom, or worth its costs. Those are further arguments to examine.

The four-part framework travels with the proposed use. The score argument asks what an output means and how uncertainty and fairness have been examined. The report argument asks how users interpret it and what actions follow. The setting argument asks whether the inference holds across learners and contexts. The public-value argument examines benefits, burdens, authority, and opportunities to challenge consequential decisions. Mislevy's account of reasoning from uncertain evidence applies across changes in assessment technology (2013, Section 6).

MAST is not an AI case study, and this case establishes no AI-related improvement in teaching or learning. It offers a concrete example of the reasoning that any proposed AI extension would need to make inspectable. Faster inference would increase the importance of knowing which claims the evidence supports.

What the MAST case contributes

MAST has put a through-year assessment into operation and produced a technical record that bears on its summative interpretations. Report research and implementation experience contribute a different body of evidence: promising features, interpretation problems, and variation in how information enters instructional decisions. These findings support different judgments about different claims. They also identify where design changes, clearer limits on use, or further studies are needed.

The next studies should address consequential uncertainties: how instructional sequence affects interpretation, what reports add to educators' existing evidence, which supports make appropriate responses possible, and whether the benefits justify the demands across different settings. Evidence of a problem may require action now; the continuing nature of validation is no reason to postpone it. Equally, an open research question is not itself evidence that a design has failed.

The field does not need early consensus. It needs claims specific enough that evidence can change someone's mind. MAST contributes to that work by making the relationship among design, evidence, use, and public judgment available for examination and revision.

References

- Badrinarayan, A. (2024). *Design principles for instructionally relevant assessment systems*. Learning Policy Institute. <https://doi.org/10.54300/111.528>
- Badrinarayan, A., Rose, C., & Fortier, E. (2026, June). *Moving toward instructional relevance: Recommendations for through-year assessment systems that advance teaching and learning*. Education First. <https://www.education-first.com/insights/our-publications/moving-toward-instructional-relevance/>
- Bell, C. A., & Mislevy, R. J. (2022). 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>
- Chu, D. (2023). *State tests that are useful for both instruction and accountability: A conversation with New Meridian's Arthur VanderVeen (Part I)*. Assessment HQ. <https://assessmenthq.org/state-tests-that-are-useful-for-both-instruction-and-accountability-a-conversation-with-new-meridians-arthur-vanderveen-part-i/>
- Chu, D. (2026, June 11). *The assessment that does everything is too good to be true*. Thomas B. Fordham Institute. <https://fordhaminstitute.org/national/commentary/assessment-does-everything-too-good-be-true>
- Dudczak, C. A., Day, D. L., & Hartwell, T. L. (1992, April 30–May 3). *The elements of CEDA debate paradigms: An investigation of paradigm accuracy* [Paper presentation]. Annual Meeting of the Eastern Communication Association, Portland, ME. <https://eric.ed.gov/?id=ED345306>
- Freeley, A. J. (1981). Judging paradigms: The impact of the critic on argument. In G. Ziegelmüller & J. Rhodes (Eds.), *Dimensions of argument: Proceedings of the second summer conference on argumentation* (pp. 433–447). Speech Communication Association. <https://files.eric.ed.gov/fulltext/ED207119.pdf>
- Huff, K. (2025). Designing and developing educational assessments for contemporary needs. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning* (pp. 131–141). University of Massachusetts Amherst Libraries. https://www.assessment-for-learning.org/files/ugd/e901ef_3a57f89efb3841f1ae676099e0b35b5d.pdf
- Ismail, K., Lamba, S., Fortier, E., & Powell, D. (2023, October). *A pilot year in review: What have we learned about through-year assessments?* Education First. <https://www.education-first.com/a-pilot-year-in-review/>
- Levitan, S., & Munyan-Penney, N. (2024, December). *5 things every equity advocate should know about through-year assessments*. EdTrust. <https://edtrust.org/wp-content/uploads/2024/11/Through-Year-Assessments-5Things-V5.pdf>
- Lichtman, A. J., & Rohrer, D. M. (1982). Policy dispute and paradigm evaluation: A response to Rowland. *Journal of the American Forensic Association*, 18(3), 145–150. <https://doi.org/10.1080/00028533.1982.11951213>
- Mislevy, R. J. (2013). *Four metaphors we need to understand assessment*. Gordon Commission on the Future of Assessment in Education / Educational Testing Service. https://www.ets.org/Media/Research/pdf/mislevy_four_metaphors_understand_assessment.pdf

References

- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
<https://doi.org/10.4324/9781315871691>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
https://doi.org/10.1207/S15366359MEA0101_02
- Montana Office of Public Instruction. (n.d.-a). *Montana aligned to standards through-year*.
<https://opi.mt.gov/MAST>
- Montana Office of Public Instruction. (n.d.-b). *Montana aligned to standards through-year FAQ*. <https://opi.mt.gov/Leadership/Assessment-Accountability/MontCAS/Required-Assessments/Montana-Aligned-to-Standards-Through-Year-FAQ>
- Montana Office of Public Instruction. (2026, March 30). *MAST memo: March 2026, Vol. 5*.
<https://content.govdelivery.com/accounts/MTOPI/bulletins/41051d3>
- Montana Office of Public Instruction, & New Meridian. (2025). *Montana aligned to standards through-year (MAST) assessment: Summative technical report, 2024–2025 school year, grades 3–8 English language arts and mathematics*. <https://opi.mt.gov/Portals/182/Page%20Files/Statewide%20Testing/MAST/MAST%202024-2025%20Technical%20Report.pdf>
- Olson, L. (2024, May). *The limitations of through-year assessments: Progress report on testing innovation*. FutureEd, Georgetown University McCourt School of Public Policy.
<https://www.future-ed.org/wp-content/uploads/2024/05/The-Limitations-of-Through-Year-Assessments.pdf>
- Park, M. J. (2024, November 6). *Through-year assessment: The importance of score reports*. Center for Assessment. <https://www.nciea.org/blog/importance-of-score-reports-for-through-year-assessments/>
- Rowland, R. C. (1982). Standards for paradigm evaluation. *Journal of the American Forensic Association*, 18(3), 133–140. <https://doi.org/10.1080/00028533.1982.11951211>
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.
- VanderVeen, A. (2026, July 9). *State assessments can be useful for both instruction and accountability—if we design them that way*. Thomas B. Fordham Institute. <https://fordhaminstitute.org/national/commentary/state-assessments-can-be-useful-both-instruction-and-accountability-if-we>