EDSTHETIC · WRITEIQ

Marking Once, Honestly: The Measurement Architecture of Writeiq

What Writeiq does for student writing, how it is built and tested, how it is calibrated against independent human judgment, and what it deliberately does not claim.

TECHNICAL WHITE PAPER · SECOND PUBLIC EDITION, AUGUST 2026

> Two experienced English teachers mark the same essay. One awards it 34 out of 50. The other awards 41. Neither is wrong. On double-rated reference sets, trained markers agree exactly on a criterion score only about half the time. Every claim made by any assessment product lives in the shadow of that fact, and most vendor accuracy pages are written hoping you will not ask about it. This paper is written assuming you will.

How to read this paper

School leaders deciding whether Writeiq can carry a whole-school literacy strategy: read sections 1 to 8 and the closing questions. About twenty-five minutes. Heads of English will find the marking and feedback machinery in sections 2 to 5. Psychometricians and technical reviewers will want sections 9 to 13, and the underlying registers, model records and test suites are available to reviewers under a confidentiality agreement. We would rather you tried to break it than took our word for it.

Where the paper uses a technical term, a short note in italics follows, beginning "what this means". The technical sentence is the claim. The note is what it means for a school.

Executive summary

Writeiq is a writing assessment platform for Years 3 to 12. A teacher sets a task once. Within minutes of students submitting, each student has feedback written for their level with one clear next step, the teacher has a marked set with the evidence highlighted and a ready-made teaching sequence, and leadership can see the whole school's writing on one screen. The teacher confirms every result, and the confirmed result is the record.

Writeiq is a writing measurement system that uses artificial intelligence, not an artificial intelligence that happens to mark writing. Underneath that sits the decision this paper keeps

returning to, because it is the single most important thing we did: the assessment framework was authored first, by people, and the artificial intelligence is only permitted to apply it. IWAF, the Integrated Writing Assessment Framework, defines every criterion, every expectation and every non-negotiable rule before a model ever sees a script. Deterministic software then re-checks the model's work after it answers, and in the secondary years the judgement itself is deliberately blind to the student's year level: the writing is assessed against the stable framework, and the approved year-level interpretation is applied afterwards, by a separate, versioned measurement layer (section 3 explains both halves). Many contemporary marking products start from the opposite end: a general-purpose model is asked to interpret a rubric supplied at runtime, and the product is built around its answers. The difference decides everything downstream, and section 2 explains why it cannot be shortcut.

On the evidence: across established public benchmarks the engine agrees with expert human markers at a level inside the published range for human markers themselves, on held-out slices it had never seen. Writeiq publishes no year-level accuracy claim, and section 10 sets out why. Sections 9 to 12 give the numbers, the method, and the limits.

The paper's evidence falls into three classes, reported separately and never substituted for one another: instrument evidence (international benchmarks, provider invariance, stress tests), Australian scale evidence (the calibration study against independently scored Australian writing), and operational transportability evidence (the prospective blind-marking pairs now accumulating under a pre-registered protocol).

1. The problem: writing assessment is slow, and humans disagree

Extended writing is the highest-signal, highest-cost assessment a school runs. The cost is familiar. A class set of essays is an evening. A year level is a week, and by the time results land the teaching moment has often passed. The signal side is discussed less. On double-rated reference sets, two trained markers agree exactly on only about half of criterion judgments. That is the published finding across the essay-marking literature, and it is why serious examination programs pay for second markers and adjudication (Shermis and Burstein, 2013; Williamson, Xi and Breyer, 2012).

Two things follow. First, an automated system should be judged against the human ceiling, not against perfection. The useful question is whether the machine agrees with a human about as well as a second human would. Researchers at the Educational Testing Service, the American organisation behind examinations such as the GRE and TOEFL, proposed exactly that standard. Second, consistency is itself valuable. The same rubric applied the same way to every script in a cohort is something human panels only approximate, through expensive moderation.

Writeiq does not replace the marker. It applies an explicit, authored framework consistently at scale, shows the evidence behind every judgment, and leaves the final word with the teacher.

2. The framework came first, and that is the whole point

Every product in this category faces a choice at its foundation. Either you hand student writing to a model and ask what it thinks, dressing the answer in a rubric afterwards, or you author the entire measurement instrument first and reduce the model to the job of applying it. The first path is common across the current generation of these products because it is weeks of work rather than years. Writeiq took the second. This section explains what was actually built, because the construction is the product's deepest moat.

IWAF defines seven dimensions of writing: Voice, Structure, Cohesion, Vocabulary, Sentence Craft, Text Structure and Conventions. Depending on the genre these expand to as many as eleven marked criteria. A narrative adds characterisation and narrative structure. A persuasive piece adds persuasive devices. Each criterion carries authored descriptors for every score point. The framework is authored in two families: a primary-years construct and a secondary construct, each written in its own developmental voice, and the writing mode selects the criterion set. The descriptors were written and refined by a literacy team against real student writing, drawing on systemic functional linguistics and the Sydney School genre tradition (Halliday and Matthiessen, 2014; Rose and Martin, 2012), so Text Structure means narrative arc for a story and claim-evidence-reasoning for an argument, never generic shape.

How year level enters the measurement deserves its own paragraph, because it is one of the most deliberate decisions in the current architecture. In the secondary years, the model that judges the writing is never told the student's year. It assesses the piece against the stable secondary construct, producing a raw criterion profile that depends only on the writing. The approved year-level interpretation — what that evidence means for a Year 7 or a Year 9 cohort — is applied afterwards, by a separate calibration layer that is versioned, frozen and applied in ordinary deterministic code. Two things follow. The judgement cannot be quietly biased by expectations attached to a year label, and the interpretation layer can be governed, audited and improved without ever touching the judgement. A student's criterion scores are the raw evidence; the overall result is the authorised interpretation of that evidence, and the product presents them that way.

Why does authorship matter so much? Three reasons.

Traceability. Every score the engine produces must cite evidence spans in the student's own writing against a named criterion with published wording. There is no "the model felt it was a B". A teacher can open any result and see the sentence that earned the judgment.

Stability. A model asked for its opinion gives a slightly different opinion each time, and a very different one each model generation. A model applying a fixed instrument, with its randomness turned off, marked once and the result stored, with a frozen calibration layer on top, gives the same essay the same treatment. Consistency across a cohort is the thing schools moderate for, and here it is structural.

Improvability. When a judgment is wrong, we can find the criterion, the descriptor or the rule that produced it and fix that one thing. A product built on model opinion can only re-prompt and hope. This is also why the construct survives model upgrades: when the underlying model improves, the instrument stays the same and simply gets applied more faithfully. Two different providers' models already score it equivalently, evidence that the measured construct attaches to the instrument rather than to the idiosyncrasies of any single foundation model.

None of this can be shortcut by a competitor with a clever prompt. The descriptors, the two authored family constructs, the rule table and the feedback library in section 4 are authored intellectual property, refined against real marking, and the calibration evidence in section 10 attaches to this instrument, not to a model that will be superseded next year.

3. Deterministic rules, and the rules we removed

A rubric says what quality looks like. Rules say what is not negotiable. Writeiq's rule system has two layers, and the second is a discipline most products never impose on themselves.

The first layer is enforcement. Technical invariants — a score can never exceed a criterion's stated maximum, can never be negative, and every criterion in the set must be scored, no more and no fewer — are checked by ordinary deterministic code after the model returns, treating the model's output as a first draft and the rule code as the authority. A language model asked to apply arithmetic constraints will occasionally miss one. Even a one percent miss rate, across thirty essays a class and hundreds of classes, is a steady stream of wrong scores. The enforcement layer takes that class of error to zero, and every applied rule is recorded on the result. The build system runs a single canonical rule table from which both the marking prompt and the enforcement code are generated, with a test suite that fails if they ever drift apart — a prompt rule the code does not enforce, or code the prompt does not disclose, cannot ship.

The second layer is subtraction, and it is the part worth reading twice. The framework's earlier generations carried a larger battery of heuristic gates — length-conditioned score ceilings, genre caps and similar rules supplied to the model as marking instructions. When we instrumented them properly, measurement showed the model applied several of these inconsistently between identical runs, and that this self-application, not the writing, accounted for real score movement. The current production rule table therefore classifies every rule: technical invariants are enforced in code; authored, evidence-validated standards are enforced in code; and heuristics that had never been validated were removed outright — from the prompt and from the code — rather than left to influence marks unaudited. A disabled rule contributes nothing anywhere, and the register records why. Rules earn their way back only with validation evidence.

What this means: the non-negotiables are checked by ordinary software that audits the AI's work, not by hoping the AI remembered. And where an old rule could not be shown to make marking fairer, it was deleted rather than trusted. The rule set schools use today is the audited one, and its record says exactly what is enforced and what was removed.

4. The permutation engine: how one marking event becomes teaching

This is the machinery the whole product exists for, and it deserves to be spelled out.

When a piece is marked, the engine produces up to eleven criterion scores, each placed in one of four achievement bands, with evidence spans and any applied rules. That profile then indexes into an authored feedback library: content written per criterion, per band, per writing mode, with primary and secondary years written separately, plus authored framing language for the rules. Four bands across up to eleven criteria across six modes, doubled for primary and secondary voices and extended by the rule framings, yields thousands of authored pathways. Two students with the same total score routinely receive different feedback, because their profiles differ, and that is the point. The feedback is assembled from human-written teaching content selected by the student's actual profile, not composed fresh by a model each time. It cannot hallucinate, it never leaks marking jargon to a child, and its tone was decided by educators.

Student One Step. Each student's feedback surfaces one focus, not a wall of advice. The engine selects the criterion where the student sits lowest relative to their band trajectory, phrases it in the band-appropriate voice, and anchors it to a highlighted example from the student's own writing. A Consolidating writer is pushed on resolving narrative tension. An Emerging writer is helped to finish a sentence. Research on formative feedback is blunt that one actionable focus beats ten observations, and the product enforces that discipline structurally.

Teach Next. The same profiles aggregate across the class, and the teacher's home screen answers what we call the Monday question: which single criterion, taught to this class next week, lifts the most students. The panel names the criterion, shows how many students sit below the class average on it, and generates a five-lesson Gradual Release of Responsibility sequence on demand: L1 teacher modelling through L5 independent practice, with teacher moves, student tasks and export to PowerPoint (Pearson and Gallagher, 1983; Rosenshine, 2012). Marking on Friday produces Monday's lesson, from evidence, in one tap.

Three audiences, one event, no repetition. The student sees their step. The teacher sees the diagnostic picture and the lesson. Leadership sees the aggregate. A hard rule with banned-phrase lists keeps the registers separate, so pipeline caveats reach the teacher panel and never a family report.

5. Seeing the whole school, and getting ready for NAPLAN

Because every class marks on the same instrument, the school-level picture assembles itself.

The leadership heatmap shows every class and every criterion on one page, updating each time a teacher confirms marks. Students flagged at risk after Emerging results on two or more tasks. Boundary students identified one mark from the next band, where a term of targeted teaching pays off most. Teaching cycles show before-and-after evidence of whether an intervention moved the

cohort. Growth lines track each student across years on the same criterion scales, feeding portfolio views and family reports that show what improved, in plain language.

NAPLAN readiness has its own view. It reports the cohort's IWAF band distribution ahead of the national assessment window, against the school's own writing evidence. It carries no year-level accuracy claim (section 10) and is not a predicted result. Writeiq is independent of ACARA, does not mark using the NAPLAN rubric and does not predict NAPLAN scores; the readiness view is a comparison against the school's own writing evidence, not a forecast of the national result. Leaders see the distribution, the movement since the last assessment, and which criteria are dragging it, while there is still time to teach. Schools using the platform's practice tasks get that signal from writing students were doing anyway, instead of from a purchased mock test.

Handwritten work participates fully. A phone photo of a page is transcribed with confidence scoring, the teacher checks the transcript, and the piece is marked through the identical pipeline. Half of Australian students still write by hand, and they are assessed on the same instrument, not a lesser one.

6. Bands, MTSS and NCCD

IWAF reports four achievement bands over the criterion-referenced result: Emerging, 0 to 30 percent. Developing, 31 to 55. Consolidating, 56 to 79. Extending, 80 and above. The 80 percent line was not chosen for neatness. It draws on the mastery-learning tradition running from Bloom through Rosenshine, the same evidence base the Australian Education Research Organisation's teaching guidance draws on.

The band architecture is designed to align with the multi-tiered system of supports, the intervention framework AERO recommends and documents for Australian schools. Extending and Consolidating together correspond to the roughly 80 percent of students well served by classroom teaching at Tier 1. Developing is a Tier 2 signal. Emerging is a Tier 3 signal. Two points of precision belong here, because being exact about endorsement is part of this paper's contract. AERO recommends MTSS as a framework. It does not endorse Writeiq or any vendor product, and the alignment claim is ours, made by design and stated as such. And Writeiq does not make tier placement decisions. Those belong to school intervention teams weighing multiple sources of evidence, of which Writeiq contributes one consistent, criterion-level strand.

The same evidence stream serves NCCD documentation. The NCCD accepts dated records of ongoing learning needs and monitored response to intervention. Writeiq produces exactly that shape of record, criterion-level and dated, with teacher adjustment notes, and it never diagnoses disability or determines a category. Those remain professional judgments.

What this means: the bands tell a teacher who needs core teaching, who needs small-group support and who needs intensive help, in the same language as the national intervention

guidance, and the paper trail is the shape the disability funding census accepts. The decisions stay with people.

7. The teacher holds the pen: confirmation, blind marking, moderation

Every number in this paper is advisory until a teacher confirms it. That is not a disclaimer. It is the operating model, and three features make it real.

Confirmation and adjustment. When a submission opens, the engine's suggested scores are labelled as suggestions. The teacher confirms, adjusts any criterion, or attaches a note where they see it differently. The original engine score is preserved unchanged for audit, every adjustment is logged with its author, and the confirmed score is the official record. Senior-year adjustments require a written rationale.

Blind marking. A teacher can flip one switch and mark the other way around: the engine's scores stay hidden, the teacher enters their own judgment first, and only after saving do they see the engine's view beside theirs, criterion by criterion, with agreement shown. This matters for two reasons. It protects professional judgment from anchoring, the well-documented pull toward a number you have already seen (Tversky and Kahneman, 1974). And every blind-marked piece is recorded as an agreement pair uncontaminated by anchoring: the raw material of the pre-registered multi-school study. A product confident in its measurement can show you your score first, and keep count honestly.

Moderation. Because every class marks on the same instrument, a year level's results are comparable by construction, and the platform supports the moderation practices schools already run: side-by-side comparison of teacher-confirmed and engine results, blind cross-marking rounds where colleagues score the same piece without seeing each other's judgments, and agreement summaries that make the conversation about evidence instead of memory. Consistency across markers stops being an aspiration managed by meetings and becomes something the school can inspect.

8. Is it safe? The questions schools actually ask

Fear of AI in schools is rational, and it deserves specific answers, not reassurance. Here are the five questions we are asked most, with the mechanism that answers each.

Will it replace the teacher's judgment? No, structurally. Suggestions are labelled as suggestions on every surface, confirmation is the teacher's, blind marking lets judgment come first entirely, and the confirmed score is the only official one. The product's teaching outputs, lessons and next steps exist to give teachers time and information, and the classroom remains theirs.

Where does the writing go? Student writing is processed and stored inside Australia, on infrastructure in Sydney and Melbourne. The residency rule is enforced in the software itself: a request that cannot be served onshore queues until it can, rather than being routed overseas. Contract terms prohibit student work being used to train external models. The full data handling policy, sub-processor list, incident response plan and deletion policy are published on the website.

What if it gets something wrong? The rule layer re-checks the model's work in deterministic code. Consistency checks flag pieces where the criterion profile and overall result disagree, routing them to human review instead of letting them pass. The teacher sees the evidence for every judgment and can override any of it. And because results are marked once and cached, an essay does not get a different mark on a different day.

Is it fair to every student? A paired test found no penalty for Australian spelling conventions against American ones. Handwritten work receives full parity through the transcription pipeline. Feedback voices are matched to achievement level so an emerging writer is taught, not graded and left. Fairness testing is ongoing and its results are reported in the evidence pack, including the tests that have not been run yet.

Who is accountable? Every score traces to named criteria, cited evidence and recorded rules. Every adjustment is logged. If a student's writing raises a wellbeing concern, safeguarding detection alerts staff first. And the company's own claims are governed by registers enforced in the build system, described in section 13, so the product is prevented from describing itself more generously than its evidence allows.

This posture maps directly onto the Australian Framework for Generative AI in Schools, approved by all Education Ministers in October 2023 and in effect since Term 1 2024. Its six principles are teaching and learning, human and social wellbeing, transparency, fairness, accountability, and privacy, security and safety. The mechanisms above answer them point by point, and the product's design assumed from day one that schools would be required to ask. State education departments increasingly run approved-tool assessments covering exactly these dimensions, and Writeiq's published documentation exists to pass them.

What this means: the national rules ask whether the teacher stays in charge, whether families can know what is happening, whether data stays onshore and safe, and whether anyone is accountable. Each has a specific mechanism here, in the software, not a paragraph in a policy binder.

9. Sixty years of machines marking writing

Automated essay scoring is older than most people expect, and its history explains why Writeiq was built the way it was.

Project Essay Grade (Page, 1966) showed that surface features like essay length could predict human scores surprisingly well, which defined the field's permanent worry: a system can correlate

with markers without measuring writing. The Educational Testing Service's e-rater system brought linguistic features and real measurement governance, and still co-scores that organisation's major examinations alongside humans (Attali and Burstein, 2006). IntelliMetric and the Intelligent Essay Assessor commercialised statistical approaches (Rudner, Garcia and Welch, 2006; Landauer, Laham and Foltz, 2003). The 2012 ASAP benchmark tested the field openly (Shermis and Hamner, 2013), and critics demonstrated that verbose nonsense could fool feature-counting systems. The current generation of large language model markers can genuinely read meaning, which removes the old weakness and introduces a new one: fluent, confident output with no guarantee of measurement discipline, and no guarantee the same essay is treated the same way twice.

One more lesson from the market deserves its own paragraph. Most published accuracy figures for classroom writing-assessment products are teacher-acceptance rates: how often teachers accepted the suggested grade. One prominent vendor reports 81.2 percent exact agreement on 579 essays, described in its own words as teachers accepting or changing the suggested grade. The weakness is anchoring again: a marker shown a score first is pulled toward it, so acceptance rates flatter the machine. They measure persuasiveness as much as accuracy. The number a school actually needs is blind external agreement: the engine compared against marks awarded independently by people who never saw its output, on scripts it was never tuned on, with confidence intervals. Serious systems do publish validation against previously human-scored corpora, and those studies matter: one widely used scoring engine reports agreement validated on more than fifty thousand human-graded responses. What remains rare is the combination this paper commits to: evaluation restricted to scripts held out from every fitting step, confidence intervals, distribution and ordering checks, a published record of the candidate models rejected along the way, and a published refusal wherever the evidence does not yet support a claim. The next section is that refusal.

10. Year-level evidence: what Writeiq does not claim

The most useful thing a measurement product can publish is the boundary of its own evidence. Writeiq publishes **no year-level accuracy claim**, at any year level from 3 to 12.

That means no exact band-agreement percentage, no within-one-band rate, no kappa and no rank correlation at any year level. Where a year level is served, its descriptors are framework-mapped against the authored IWAF criteria, and the product labels them as such rather than implying an empirical calibration that does not exist.

The evidence Writeiq does publish is engine-level, measured against openly licensed external corpora with holistic rubrics (section 12). That evidence speaks to whether the engine discriminates writing quality consistently with expert human judgement across prompts, collections and grade bands. It does not establish a year-level calibration, and this paper does not present it as one.

Establishing year-level evidence requires a prospective, multi-school study in which teachers' own confirmed judgements form the comparison, with the participating schools' authorisation and

appropriate consent in place before any student work is used. Section 14 describes that programme. Until it reports, the honest position is refusal, and this section is the refusal.

What this means: Writeiq will not tell you how often it agrees with a year-level standard, because it has not published evidence that would make such a number meaningful.

11. Across ten year levels, without pretending

A whole-school scale cannot treat every year level as equally evidenced, so Writeiq states the standing of each. No year level carries a published accuracy claim. Years 10 and 11 are not calibrated years: they receive no calibration adjustment at all, deliberately, because the available evidence did not meet the bar and one external dataset showed a bias that made a calibration claim untenable. The honest action was refusal, and the refusal is written down. Year 12 is the authored senior anchor, a design commitment rather than an empirical claim. The method is frozen, with the conditions for reopening it published, so the scale cannot drift through quiet refitting.

What this means: where the evidence was not good enough, nothing was adjusted and the reasons are recorded. Senior results are presented as indicative and teacher-confirmed.

12. International benchmarks: testing the instrument, not borrowing it

To test whether IWAF measures writing rather than one cohort's marking culture, the framework was applied unchanged to the openly licensed ASAP 2.0 and PERSUADE research datasets and evaluated on held-out slices. The two corpora overlap substantially, so the figures below are reported on the non-overlapping partition and are one line of evidence, not two independent replications. Applied cold to American prompts with no adaptation, the locked test gave quadratic weighted kappa 0.737 on 889 held-out essays, where systems purpose-trained on those prompts reach about 0.80, and 85.9 percent of judgements fell within one score point. Stress tests behaved the way a real instrument should: agreement held when whole prompts were withheld from calibration and the held-out prompt was scored cold, 0.740 pooled across all seven, and it degraded gracefully on the hardest transfer, unseen grade levels at 0.636, which is reported anyway. The spread of the scores tracked the human spread rather than collapsing toward a safe middle band. A paired test found no penalty for Australian spelling. Two independent model providers scored equivalently.

The instrument has also been evaluated against additional research corpora held under non-commercial licences. Those results are research evidence and are not reported here; they are available to reviewers under a confidentiality agreement, alongside the registers and model records.

These results are kept separate from the Australian claims by design. They evidence the instrument. They are never used to claim anything about Australian year-level bands.

What this means: the same framework, with nothing retuned, marks American persuasive essays and English-learner writing about as consistently as those datasets' official human markers. That says the framework measures writing itself. We still do not borrow it as proof about Australian bands.

13. Trying to break it, and keeping it honest in production

A measurement claim is worth what the attempts to break it are worth. Three attempts are on the record. A fully nonparametric re-derivation of the calibration, discarding every statistical assumption of the production model, reproduced 90 percent of its decisions, with differences confined to sparse edges: the result is not an artefact of its assumptions. A leave-batch-out analysis confirmed the calibration is not an averaging artefact. And one correction is disclosed because its shape is the point: an earlier internal claim that two criteria correlated with an external measure turned out to rest on zero formed data pairs, the analysis script having silently skipped them. The claim was struck, the register corrected, and the pipeline changed to record sample sizes so silent omission cannot recur. Higher-scoring candidate models that failed the compression checks are retained in the record as rejections.

Production is engineered so the version schools use is provably the version that was tested. Every scoring configuration — the judgement engine, its prompts and rule table, and the calibration layer — belongs to a named, immutable scoring release. Calibration artefacts are versioned, approver-stamped, fail-closed and pinned byte-for-byte: the build system recomputes the live engine's identity fingerprints and the calibration artefact's checksum on every build and refuses to ship on any mismatch. That guard is not theoretical; it has already blocked one of our own deployments over a one-byte packaging discrepancy, which is exactly what it is for. Which release a school runs on is decided by a server-held, per-school authority that client software can read but never choose, with an unknown or invalid value failing safely to the incumbent release. Every mark carries its release lineage, a new release reaches a school only through a controlled, evidence-gated cutover with a rehearsed configuration-only rollback, and rolling back never re-scores or re-labels historical results: a mark keeps the release that produced it, forever. Reproduction suites prove the served values match the research artefacts exactly. The same engineering governs language. A claims register records approved and prohibited wording for every finding. Each of the sixteen curriculum framework surfaces carries a validation register with a four-step ladder from indicative to framework-validated, and today every surface sits on the first step and says so in the product. A build-time gate fails if a surface ships without its register entry, drops its indicative labelling, or uses evidence-grade language it has not earned.

What this means: the mathematics a school uses is provably the mathematics that was tested, and marketing physically cannot outrun the data, because the build fails if it tries.

14. Limits, and what happens next

Stated plainly, because a paper that hides its edges cannot be trusted about its centre. Writeiq carries **no year-level accuracy claim at any year level from 3 to 12**; building that evidence, in Australian classrooms and with the participating schools' authorisation, is the open empirical question. The international results evidence the instrument, not Australian bands. Every framework-native surface is indicative until its own register says otherwise. And teacher-acceptance figures, however flattering, are a different thing from blind accuracy; the prospective blind study that separates them cleanly is the planned next step. The thresholds for that study are pre-registered and frozen (MSBA-1, 31 July 2026); the registration's fingerprint is published on the research page, and any edit to the registration fails our build system.

The instrument for that next step is already in classrooms. Blind marking reveals the engine's view only after the teacher commits their own, so every blind-marked piece contributes an uncontaminated agreement pair, and the school-level calibration step of each framework register is reached through exactly this data. The evidence base this paper stands on is designed to grow through ordinary use.

Eight questions to ask any vendor

If this paper does one thing for a procurement process, let it be this list. These are the questions we set for ourselves, and every product in the category should be able to answer them in writing.

1. Is your headline accuracy figure blind, or did the people who set the truth marks see the AI's suggestion first? Writeiq: blind. The truth set was marked independently before the engine existed.
2. Was it measured on scripts the system was never tuned on, with confidence intervals? Yes. Locked holdout, intervals published.
3. Would your number survive a collapse check? Exact agreement can be bought by predicting the most common grade. Show spread and ordering evidence. Yes: spread matches the human markers, and the ordering test fixes 4.6 rankings for every 1 it breaks.
4. What do you refuse to claim? Section 14, with a register of prohibited wording behind it.
5. Which negative results will you show me? Section 13, including a self-reported correction.
6. Is the served product provably the validated one? Yes. Reproduction and release-identity suites run on every build, and versioned artefacts fail closed.
7. When you present my curriculum's construct, is it validated or indicative, and does the product say so on every surface? A verdict ladder per framework, enforced in the build.
8. Where does student writing go, and is that enforced in code or only in a policy document? Australian processing and storage, refusal-first fallback, tested in the build.

A vendor who can answer all eight has done the work. A vendor who answers with a single percentage has given you a number, not a measurement.

Closing

The single most important decision in Writeiq is the oldest one: the instrument was authored before the intelligence was applied, and everything else follows from it. The deterministic rules that make the non-negotiables non-negotiable, and the honesty to delete the rules that could not earn their keep. The year-blind judgement that keeps expectations out of the evidence, and the governed interpretation layer that gives the evidence its year-level meaning. The feedback library that turns a criterion profile into one next step for a child and one Monday lesson for a class. The analytics that let a leader see writing across a school, and the NAPLAN readiness signal that comes from work students were doing anyway. The blind marking that protects a teacher's judgment while quietly building the strongest evidence in the category. And a claims system that keeps the company as honest as the product.

We invite the comparison. Bring the eight questions.

References

- Attali, Y., and Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3).
- Australian Education Research Organisation (2024). *Introduction to a multi-tiered system of supports: explainer*, and *Multi-tiered system of supports: user guide*. edresearch.edu.au.
- Australian Government Department of Education (2023). *Australian Framework for Generative Artificial Intelligence (AI) in Schools*.
- Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2). Bloom, B. S. (1984). The 2 sigma problem. *Educational Researcher, 13*(6).
- Cohen, J. (1968). Weighted kappa. *Psychological Bulletin, 70*(4), 213 to 220.
- Crossley, S. A., and colleagues (2022). The PERSUADE corpus. *Assessing Writing, 54*.
- Graham, S., and Perin, D. (2007). *Writing Next: Effective strategies to improve writing of adolescents in middle and high schools*. Alliance for Excellent Education.
- Halliday, M. A. K., and Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar*, 4th edition. Routledge.
- Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*, 3rd edition. Springer.
- Landauer, T. K., Laham, D., and Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis and Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*.

- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(5), 238 to 243.
- Pearson, P. D., and Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology, 8*(3), 317 to 344.
- Rose, D., and Martin, J. R. (2012). *Learning to Write, Reading to Learn*. Equinox.
- Rosenshine, B. (2012). Principles of instruction. *American Educator, 36*(1), 12 to 19.
- Rudner, L. M., Garcia, V., and Welch, C. (2006). An evaluation of IntelliMetric essay scoring. *Journal of Technology, Learning, and Assessment, 4*(4).
- Shermis, M. D., and Burstein, J. (Eds.) (2013). *Handbook of Automated Essay Evaluation*. Routledge.
- Shermis, M. D., and Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In the *Handbook* above.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*(4157), 1124 to 1131.
- Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2 to 13.

Internal evidence sources, restricted and available to reviewers under a confidentiality agreement: the accuracy claim sheet, technical evidence report, scaling claims registers, framework criteria and validation registers, model selection and activation gate records, and the serving-integrity test suites.

Edsthetic, 2026. This paper describes the measurement architecture as at 13 August 2026. Figures are traceable to the internal evidence pack. The scope limitations in section 14 are part of the claim, not a footnote to it.