

Measuring Reading Growth: The Free Reading Standard and Microsoft Reading Progress in the Classroom

Jasper Vande Moortel · Orion-Solution · Version 0.1 (draft) · July 2026

Abstract

A child who reads 90 words per minute in October and 90 words per minute in May looks, on paper, like a child who did not grow. If the May text was two years harder, the same numbers describe substantial growth that the measurement failed to see. This paper argues that classroom reading-growth measurement systematically confounds fluency with text difficulty, quantifies the problem from the curriculum-based measurement literature (nominally equivalent passages differing by tens of words per minute; standard errors that rival a year's expected gain), and proposes a framework that separates the two: oral reading fluency measured by Microsoft Reading Progress on one axis, text difficulty measured by the Free Reading Standard on the other, joined by the assignment-to-text linking that classroom tooling has until now failed to preserve. We define difficulty-adjusted fluency, a reader's observed accuracy and rate relative to what the difficulty of the specific text predicts, and growth as the trajectory of that residual, a construct for which we find no published precedent. We describe the deployed measurement infrastructure (a Teams application that scores every assigned text on the FRS scale, links assignments to texts, and ingests per-student Reading Progress outcomes nightly, under a consent and privacy architecture described here), the Flemish norm context in which growth judgments must be framed, and a pre-registered design for the empirical study the infrastructure enables. Two honesty points anchor the paper: Reading Progress's own speech-recognition scoring has, to our knowledge, no published validity study, and the framework inherits that unknown; and the framework itself is specified before data, so that the coming classroom evidence tests it rather than decorates it.

Keywords: oral reading fluency, reading growth, curriculum-based measurement, text difficulty, Microsoft Reading Progress, progress monitoring

1. Introduction

Reading growth is the outcome teachers actually care about, and it is measured worse than almost anything else in the classroom. The dominant tradition, curriculum-based measurement of oral reading fluency, counts words read correctly per minute on short probes and tracks the count over time (Deno, 1985). The tradition's strengths are real: the measure is

quick, cheap, sensitive to change, and strongly correlated with broader reading competence. Its weakness is equally real and less advertised: the number depends on the text as much as on the child, and the tradition's answer to that dependency has been to hold text difficulty constant rather than to measure it.

Holding difficulty constant fails twice. It fails technically: passages engineered to be equivalent are not, with nominally parallel probes differing by tens of words per minute in average score, and the standard error of a single fluency measurement estimated at five to ten words per minute (Christ & Silberglitt, 2007), a range that rivals or exceeds a typical year's expected gain. And it fails pedagogically: real classrooms do not read equated probes, they read books and articles whose difficulty is supposed to rise. A growth measure that requires difficulty to stand still cannot describe the growth that matters, which is precisely the ability to hold accuracy and pace while the text gets harder.

The missing ingredient is not better fluency measurement but a second axis. Text difficulty is measurable on continuous, validated scales, and the convergence of independently constructed difficulty measures is well documented (Nelson, Perfetti, Liben, & Liben, 2012). The Free Reading Standard (FRS) provides such a scale openly, for Dutch and English, with published parameters and uncertainty bands (Vande Moortel, 2026a, 2026b). Meanwhile, classroom fluency data has quietly become abundant: Microsoft's Reading Progress, deployed through Teams in ordinary schools, records accuracy, rate, and a taxonomy of oral reading errors for every submitted reading. What neither system provides alone, and what no published work we could find provides at all, is the join: per-reading outcomes attached to the measured difficulty of the specific text that was read, accumulated per child across school years.

This paper specifies that join. Section 2 reviews the two traditions and the instrument. Section 3 defines the framework: difficulty-adjusted fluency and growth as a residual trajectory. Section 4 describes the deployed infrastructure that produces the data. Section 5 states the Flemish norm context in which growth judgments are communicated to teachers. Section 6 pre-registers the design of the empirical study the infrastructure enables. The framework is published before the data on purpose.

2. Background

2.1 Curriculum-based measurement and the passage problem

Curriculum-based measurement (CBM) established oral reading fluency as the classroom's vital sign for reading (Deno, 1985): words correct per minute (WCPM), sampled briefly and often, plotted against an aimline. Decades of use produced norm tables and decision rules, and also a sober empirical literature on the measure's error structure. Two findings matter here. First, growth is not the linear march the aimline assumes: large-sample modeling of within-

year trajectories across grades 1–8 samples shows nonlinear, decelerating growth (Nese et al., 2013). Second, and central to this paper, the passage-equivalence problem: probes constructed to be interchangeable differ substantially in difficulty, inflating the error of any single measurement to the point where Christ and Silberglitt (2007) place the standard error at five to ten WCPM. The CBM tradition treats this as noise to engineer away through passage equating. We treat it as signal that was never measured: if passages differ in difficulty, then difficulty belongs in the model, not in the error term.

2.2 Text difficulty as the second axis

The companion papers specify how FRS measures text difficulty: a deterministic linguistic feature core and a calibrated model per language, producing a continuous 0–100 score with an uncertainty band and a four-dimension profile, anchored for Dutch with a published mapping to the AVI region teachers already know (Vande Moortel, 2026a, 2026b). For this paper's purposes the essential property is not FRS's particular construction but its openness and its per-text applicability: any text a teacher assigns, including texts the teacher wrote or generated that morning, receives a difficulty measurement with known provenance. The broader literature supports treating such measures as trustworthy first-pass difficulty axes: independently developed difficulty metrics, from Lexile to grade-level formulas, correlate strongly with each other and with student performance despite entirely different internals (Nelson et al., 2012).

2.3 Microsoft Reading Progress as a fluency instrument

Reading Progress is a reading-fluency tool inside Microsoft Teams for Education: a student reads an assigned passage aloud to the camera, speech recognition scores the recording, and the teacher receives accuracy and rate plus a detailed error taxonomy. Its Graph API reporting surface exposes, per submission: accuracy score, words per minute, word count, and counts of mispronunciations, omissions, insertions, repetitions, self-corrections, and unexpected pauses, along with punctuation-related miscues and a monotone-score prosody proxy (Microsoft, 2025, `readingAssignmentSubmission` resource documentation). This is, in CBM terms, a remarkably rich automated oral reading record, produced as a byproduct of ordinary classroom practice at zero marginal assessment cost.

Two facts temper the enthusiasm, and this paper depends on stating them plainly. First, Microsoft's own published research page for the tool cites the general fluency-pedagogy literature but, as far as we could establish, no study validating the tool's speech-recognition scoring itself: how ASR-scored accuracy and WCPM relate to human-scored oral reading, for which accents, ages, and languages, is publicly unknown. Independent evaluations exist but measure downstream outcomes rather than measurement validity. Every result built on Reading Progress data therefore inherits an unquantified instrument error, and the study design of §6 includes a human-scoring calibration arm for exactly this reason. Second, Reading Progress

reports outcomes per assignment, not per text-difficulty: the tool itself has no difficulty model, which is the gap FRS fills and the reason the linking infrastructure of §4 has to exist.

2.4 What growth should mean

Combining the axes yields a definition. Let a reading event be a triple: a reader, a text with FRS difficulty d (and its uncertainty), and an outcome vector (accuracy, WCPM, error profile). Define **expected fluency at difficulty**, per language and age band, as the empirical relation between d and the outcome distribution across readers; fitting these curves from real classroom events is the §6 study's second question, and no empirical curve is claimed here (the application's synthetic demonstration data encodes a monotone decline by construction and proves nothing). Then a reader's **difficulty-adjusted fluency** is the residual of their observed outcome against that expectation, and **growth is the upward drift of the residual trajectory**, or equivalently the rightward drift of the difficulty at which the reader maintains criterion fluency. A child whose WCPM is flat across a year of rising text difficulty has a rising residual: measured growth. A child whose WCPM rises on texts of flat difficulty may or may not be growing; the residual says. We find no published metric of this form: the CBM tradition equates difficulty away, and the difficulty-scale tradition models comprehension criteria rather than fluency conditioned on difficulty. The construct is simple enough that its absence is more likely an infrastructure gap than an intellectual one, which is the point: the data to fit expected-fluency curves has not existed in linked form until now.

3. The framework, stated compactly

1. **Every assigned text is measured.** Library texts, pasted texts, and generated texts all carry an FRS score, version, uncertainty, and profile before they reach a student.
2. **Every reading outcome is linked.** The assignment-to-text link is stored at assignment creation, so each Reading Progress submission lands attached to the difficulty of what was actually read.
3. **Expected-fluency curves are fit per language and age band** from accumulated linked events, with the fitting population, exclusions, and refit cadence published, following the same batched versioned discipline as the scale itself (Vande Moortel, 2026d).
4. **A reader's state is a residual with uncertainty; growth is its trajectory.** Reports to teachers foreground the trajectory and its confidence band, never a bare number.
5. **Norm framing is local and honest.** Comparisons are presented against the Flemish reference material described in §5, labeled as expected levels where they are definitional, never as measured medians where no median was measured.

4. Measurement infrastructure

The framework is implemented in a deployed Teams application whose relevant properties are measurement properties. Texts are scored server-side at save time by the FRS service (deterministic, versioned); teacher-generated texts are scored and re-scored in a verification loop before entering the library, so generated material cannot claim an unverified level. Assignments store the text linkage at creation. A nightly ingestion pulls Reading Progress submissions through the Graph reporting endpoint and lands them against the linked text with its difficulty metadata. Student identity is durable across school years (the identity model that classroom tooling commonly breaks at year boundaries), so trajectories span years by construction. Consent and privacy are architectural but must be stated precisely: the application processes classroom data on behalf of the school as controller, with per-tenant Graph credentials and tenant-scoped, least-privilege data partitioning in the application's database; child-level records therefore reside in the application operator's infrastructure, not solely in the school tenant, and school agreements must say so. The research use of §6 will operate only on consented, pseudonymized exports; that export pipeline exists at the time of writing as specification, not yet as code, and building it is a stated precondition of the study, not an assumed feature. The infrastructure's test and verification discipline, including end-to-end journey verification against seeded multi-year data, is documented with the implementation.

5. The Flemish norm context

Growth judgments reach teachers as comparisons, and comparisons import norms. For Flanders the honest situation is specific: Dutch AVI and DMT instruments carry Dutch norms that demonstrably do not transfer (Flemish pupils score lower on DMT in all six primary grades and read AVI cards more slowly in all but one; Leysen, Van den Broeck, Keuning, Noé, & Geudens, 2018), a Flemish renorming exists for a single mid-year test moment with modest samples in the upper grades, and no published distribution of attained AVI level per grade exists at all, only the instrument's definitional expected level. The application therefore presents the Flemish DMT percentile fan and the definitional expected level, labeled as such ("verwacht niveau"), with the norming study's own validity caveats displayed rather than buried. This is a smaller promise than "compare your class to Flanders" and it is the promise the evidence supports; the framework's own accumulated data offers the long-run path to something better, namely empirical expected-fluency curves for the population actually using the system.

6. Pre-registered design for the empirical study

The study the infrastructure enables, stated before its data exists.

Questions. (Q1) Instrument: how do Reading Progress ASR scores relate to human scoring of the same recordings, per language and age band? (Q2) Curves: what are the empirical expected-fluency-at-difficulty curves, and how do they compare to the CBM norm tradition where difficulty overlaps? (Q3) Growth: do difficulty-adjusted residual trajectories detect growth that flat WCPM tracking misses, and vice versa? (Q4) Seasonality and shape: do the trajectories reproduce the nonlinear within-year patterns of the CBM literature (Nese et al., 2013)?

Design. One to three volunteer Flemish schools, one school year, ordinary classroom use (no experimental manipulation of instruction); explicit opt-in consent per family with a non-participation path that changes nothing about classroom experience; a human-scoring arm in which a stratified sample of recordings is double-scored by trained scorers for Q1. Analyses: mixed-effects growth models on residual trajectories (reader random effects, text difficulty and its uncertainty as measured covariates), with the CBM comparison run on the same events reduced to raw WCPM. Falsifiers are stated in advance: if Q1 correlations are poor, the instrument arm gates everything else; if residual trajectories show no incremental signal over raw WCPM (Q3), the framework's added complexity is not justified and we will say so.

What this paper does not claim. No growth result, no instrument validity result, no norm. This is the specification and its justification; the results paper is separate by design.

7. Discussion

The framework's intellectual content is deliberately modest: measure both axes, keep the join, define growth as conditional performance. Its practical content is larger, because the join is exactly what classroom tooling has not kept, and because each component now exists in deployed, tested form: an open difficulty scale with published uncertainty, a fluency instrument schools already use, and the linking infrastructure between them. The framework also positions the long-run psychometric ambition stated in the standard's specification: linked reader-text outcome data at scale is the prerequisite for conjoint calibration in the Rasch tradition, which would put readers and texts on one scale with comprehension-style semantics. That remains future work and is not smuggled in here.

The deepest dependency is the instrument. If Reading Progress's ASR scoring proves weakly related to human scoring for Flemish children's Dutch, much of the pipeline's cheapness evaporates into measurement error, and the honest response will be to say precisely that, with numbers. The study is designed so that this is the first thing learned, not the last.

8. Limitations

No empirical growth data yet; the framework is specified, not validated. Reading Progress ASR validity is publicly unknown and inherited. Expected-fluency curves require volume before their own uncertainty is acceptable; early reports will carry wide bands. The Flemish norm material covers one test moment with thin upper-grade samples. Several CBM-tradition citations rest on our source review's secondary reporting pending full bibliographic verification, and the negative claim "no published difficulty-adjusted fluency metric" is a bounded search result, not a proof of absence. Comprehension is out of scope by construction: this framework measures technical reading growth only.

9. Conclusion

Growth is a relation between a reader and rising difficulty, and it should be measured as one. The tools to do so at classroom scale now exist together for the first time: an open text-difficulty standard, a ubiquitous automated fluency instrument, and infrastructure that refuses to lose the link between them. What remains is the discipline this paper commits to: validate the instrument before trusting it, publish the curves before judging children against them, and let the first school year of linked data test the framework in public.

References

- Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measurement of oral reading fluency. *School Psychology Review*, 36(1), 130–146. [Details via secondary reporting; pending primary verification.]
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52(3), 219–232. [Details via secondary reporting; pending primary verification.]
- Leysen, H., Van den Broeck, W., Keuning, J., Noé, M., & Geudens, A. (2018). *Vlaamse normering van de Drie-Minuten-Toets en AVI-toetskaarten van 2009*. Thomas More Antwerpen.
- Microsoft. (n.d.). *readingAssignmentSubmission resource type* (Microsoft Graph v1.0 documentation). Microsoft Learn. Retrieved 2026-07-07. [Primary API documentation, fetched directly; the production ingestion uses the beta reporting endpoint, whose page awaits its own source capture.]
- Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Student Achievement Partners. [Findings cited via secondary summaries; full report pending primary verification.]
- Nese, J. F. T., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across Grades 1–8. *Journal of School Psychology*, 51(5), 625–642. [Details via abstract-level reading; pending full primary verification.]

Vande Moortel, J. (2026a). *The Free Reading Standard: Definition and construction of an open, continuous scale for technical reading difficulty in Dutch and English* (v0.2). Orion-Solution. <https://frs.orion-solution.net/research>

Vande Moortel, J. (2026b). *Calibrating an open reading-difficulty scale without human-labeled corpora* (v0.1). Orion-Solution. <https://frs.orion-solution.net/research>

Vande Moortel, J. (2026d). *A living standard: Crowdsourced comparative judgment and batched recalibration of an open reading-difficulty scale* (v0.2). Orion-Solution. <https://frs.orion-solution.net/research>

Entries whose bibliographic details rest partly on secondary sources are marked; they are resolved before any published version.

Change notes v0.1: first draft, framework and pre-registered design; no empirical results by design. Figures planned for v0.2: the two-axis framework, an expected-fluency-at-difficulty illustration from seeded demo data (clearly labeled synthetic), the infrastructure pipeline, and the residual-trajectory report mock.