

Estimating Reader Ability on an Open Text-Difficulty Scale from Oral Reading Performance: A Conjoint Model with Simulated Uncertainty Behavior

Jasper Vande Moortel, Orion-Solution CommV Version 0.1 — 10 July 2026

Abstract

When a reader's ability and a text's difficulty are expressed on the same scale, one number answers the question teachers actually ask: which texts can this reader handle. The Lexile Framework operates such a conjoint pair for oral reading in production, but its estimation model is closed and its uncertainty behavior after one or a few readings is unpublished. This paper specifies an open conjoint model that places reader ability on the FRS 0 to 100 text-difficulty scale from oral reading observations (words correct per minute and accuracy on texts of known difficulty), publishes every constant with its provenance, and derives the model's uncertainty behavior by simulation, since we could locate no published empirical reference for it. Three results carry the paper. First, the estimator is statistically well calibrated: across 225 simulation cells its nominal 95 percent bands cover the true ability 93.8 to 96.4 percent of the time, including after a single reading. Second, the single-reading band is approximately 10.5 FRS points wide at baseline constants and narrows about 40 percent by the third reading. Third, the dominant driver of uncertainty is not measurement noise but the provisional mapping between the FRS scale and the Lexile-derived slope constants: varying that one assumption by ± 50 percent swings the single-reading band from 7.2 to 18.1 points, larger than observation noise or text choice. The model is calibration-limited, not noise-limited, defining what empirical reading data must buy. All constants ship in a versioned artifact designed to be replaced, not patched, when that data arrives.

1. Introduction

A fluency score describes a reading; teachers need a statement about a reader. The gap between the two is an estimation problem: given one or a few oral readings of texts whose difficulty is known, locate the reader on the difficulty scale, with honest uncertainty. When reader ability and text difficulty share a scale, the resulting statement is directly actionable ("reads texts up to about FRS 45 fluently"), and the scale does double duty: the same number that calibrates a library calibrates its readers.

This conjoint design has one production-scale precedent. MetaMetrics operates a paired Lexile oral reading measure (the student) and Lexile oral readability measure (the text) on the common Lexile scale, estimating the student measure from words correct per minute (WCPM) and accuracy on passages of known oral readability (MetaMetrics, n.d.). The design descends from the Rasch tradition that built the original Lexile scale (Stenner, 1996). But the production model is closed: the estimation procedure is not published beyond two slope coefficients and a worked example, and, critically for classroom use, nothing is published about its error behavior. How wide is the ability estimate after one reading? How fast does it narrow? Under what text-choice conditions does it degrade? Those questions decide what a responsible user interface may display, and to our knowledge no answers are published for any production system.

This paper answers them for an open model. It contributes:

1. **An open conjoint estimation model** on the FRS 0 to 100 scale (Vande Moortel, 2026d): expected-performance functions for WCPM and accuracy given reader ability and text difficulty, and a maximum-likelihood estimator with analytic uncertainty. Every constant is published with its provenance (Section 3, Appendix A).
2. **Simulated uncertainty behavior**: standard-error-versus-number-of-readings curves, coverage calibration, and text-choice effects (Section 4), filling the gap the production precedent leaves open.
3. **A sensitivity analysis** identifying the scale-mapping assumption, not observation noise, as the dominant uncertainty driver, which converts a vague "constants are provisional" caveat into a quantified statement of what empirical calibration is worth (Sections 4.3 and 5).
4. **Binding interface rules** derived from the simulation, for the public deployment of this model in the FRS Reading Check (Section 5).

The model consumes the per-reading measurements defined and validated in the fifth paper of this series (Vande Moortel, 2026e); its empirical calibration on real readings is registered follow-up work, not claimed here. Everything in this paper is a specification plus its simulated behavior; we state that plainly and design for it, shipping the constants in a versioned artifact (reader-vo) whose replacement by empirically fitted values is the intended lifecycle, mirroring the batched recalibration doctrine of the FRS scale itself (Vande Moortel, 2026a).

2. Background

2.1 The conjoint precedent

The Lexile oral reading system pairs a student measure and a text measure on one scale. Its published mechanics amount to: at 95 percent accuracy, a 100L increase in student ability

corresponds to about +14 WCPM, and a 100L increase in text difficulty to about -9 WCPM; a worked example (a 600L reader on an 800L text reading 77 WCPM) anchors the intercept; and the text-side oral readability measure uses a six-feature formula that differs from the silent-reading Lexile formula, evidence that oral reading difficulty is a related but distinct construct (MetaMetrics, n.d.). The system was calibrated on more than 3,000 passages spanning six commercial oral reading fluency assessments. What is not published: the estimation procedure that turns observations into a student measure, the treatment of accuracy, the minimum-readings guidance, and any standard-error behavior.

2.2 What a reading observation is worth

The estimation problem's noise floor comes from the curriculum-based measurement literature: the standard error of measurement for a single WCPM observation is on the order of 5 to 10 words correct per minute, driven substantially by passage-to-passage variation even within controlled probe sets (Christ & Silberglitt, 2007). Norms give the expected magnitudes the model must reproduce: median WCPM roughly doubles between grade 2 and grade 5 in the compiled United States norms (Hasbrouck & Tindal, 2017). Dutch ASR-based fluency measurement adds the instrument's own error on top; the FRS meter's instrument characterization and mechanics validation are reported separately (Vande Moortel, 2026e), and recent Dutch validations of ASR-derived fluency measures give the external context (Harmsen, van Hout, Cucchiaroni, & Strik, 2025; van der Velde et al., 2025).

2.3 Why not item response theory with binary items

The orthodox psychometric treatment would model each passage as an item and fit a Rasch or IRT response model. We considered and rejected it for this setting, for three reasons. First, the natural observations here are continuous (WCPM, accuracy), and dichotomizing them ("read fluently: yes/no") discards most of their information precisely when data is scarcest. Second, the anonymous public deployment yields one to three readings per user; binary-response ability estimation is degenerate at those counts (a perfect or zero score yields an unbounded maximum-likelihood estimate), while continuous observations keep the likelihood well behaved from the first reading. Third, the production precedent facing the same constraints made the same choice: the Lexile oral reading student measure is a regression-style estimate from continuous rate and accuracy, not a binary-item IRT fit (MetaMetrics, n.d.). The Rasch tradition still supplies the conjoint framing (Stenner, 1996); the response model is where we depart.

3. The model

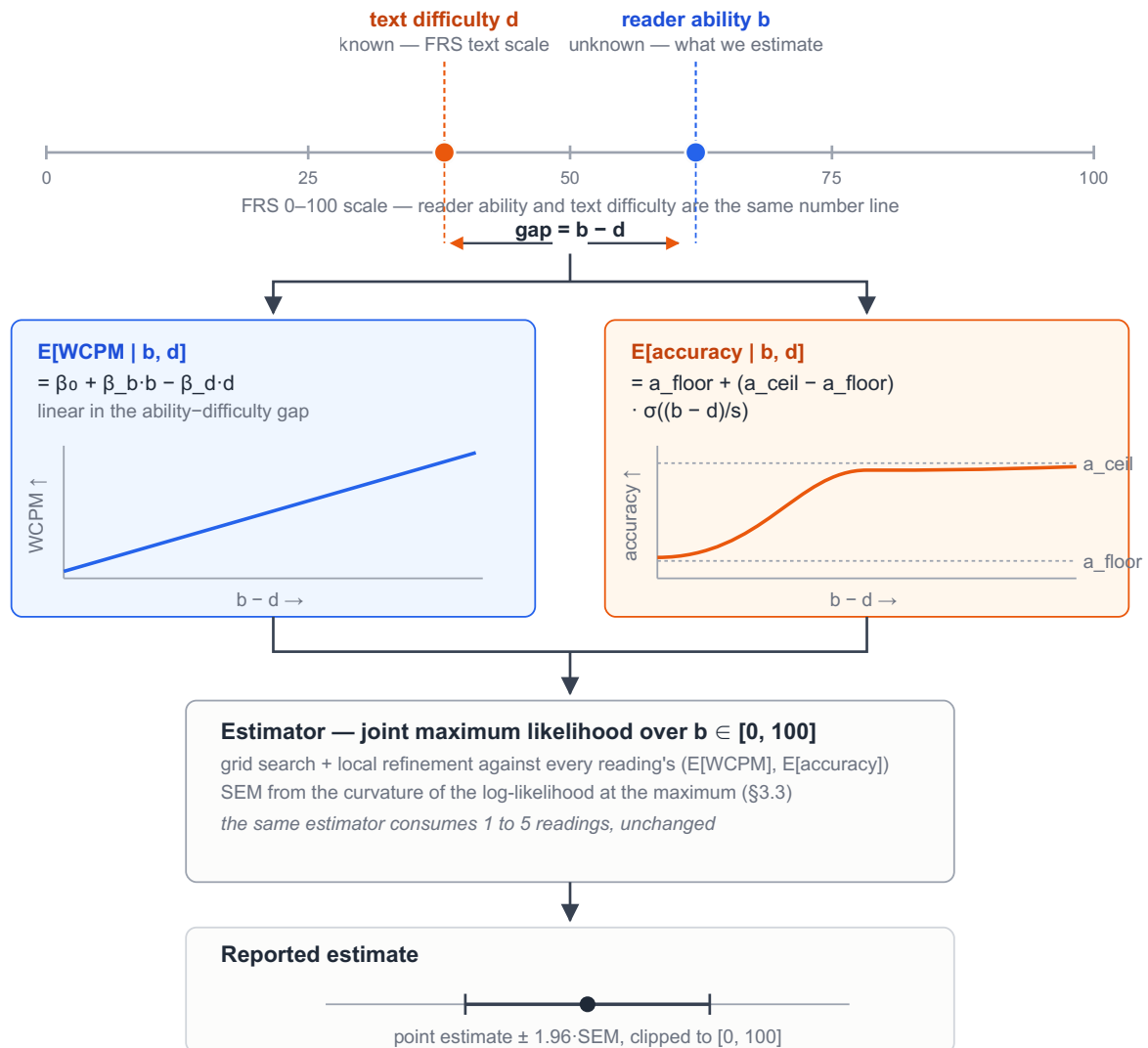
3.1 Expected performance

Reader ability b and text difficulty d share the FRS 0 to 100 scale (Figure 1). The vo model uses a linear rate function and a logistic accuracy function:

- $E[\text{WCPM} \mid b, d] = \beta_0 + \beta_b \cdot b - \beta_d \cdot d$
- $E[\text{accuracy} \mid b, d] = a_{\text{floor}} + (a_{\text{ceil}} - a_{\text{floor}}) \cdot \sigma((b - d) / s)$

with σ the standard logistic function. Observed WCPM is modeled as Gaussian around its expectation with standard deviation σ_{wcpm} . Observed accuracy enters through a binomial log-likelihood: the reading is treated as n_{eff} independent word-level Bernoulli trials with success probability $E[\text{accuracy} \mid b, d]$, with $n_{\text{eff}} = 150$ as a fixed effective word count (a modeling constant, since the anonymous deployment's schema does not carry per-reading word counts; it is published in the artifact like every other constant and is replaced by the true count wherever one is available).

Figure 1 — The conjoint model: reader ability and text difficulty share one scale



Text difficulty d is known from the FRS text scale; reader ability b is unknown and is what the estimator locates (§3.1). Every reading feeds both functions into the estimator; band-width behavior is derived by simulation, not assumed (§4).

3.2 Constants and their provenance

Every constant ships in the versioned artifact `reader_model_v0.json` with a provenance record. The derivation chain, in full:

- The span assumption.** The published Lexile slopes are per 100L. To express them per FRS point requires a mapping between the scales. No empirical mapping exists yet (that is what calibration is for), so *vo* adopts a stated assumption: FRS 0 to 100 spans approximately the Lexile range 0 to 2000L, that is, 20L per FRS point. This is an assumption, recorded as such, and Section 4.3 quantifies exactly how much rides on it.
- Slopes.** Under that assumption, +14 WCPM per 100L of ability becomes $\beta_b = 2.8$ WCPM per FRS point, and -9 WCPM per 100L of difficulty becomes $\beta_d = 1.8$ WCPM per FRS point (MetaMetrics, n.d.).
- Intercept.** $\beta_0 = 65.0$ WCPM, solved from the published worked example (600L reader, 800L text, 77 WCPM) under the same mapping.

4. **Accuracy function.** $a_{\text{floor}} = 0.85$, $a_{\text{ceil}} = 0.99$, $s = 12$ FRS points, anchored so that the ability-equals-difficulty point and the extremes reproduce the informal-reading-inventory convention bands (independent roughly ≥ 97 percent, frustration below roughly 90 percent), whose threshold values are themselves conventions with documented variation rather than empirical constants.
5. **Observation noise.** σ_{wcpm} is not fixed; the simulation sweeps 5, 8, and 10 WCPM, the passage-equivalence range reported for curriculum-based measurement (Christ & Silberglitt, 2007).

Sanity anchoring: with these constants, simulated WCPM values across the ability grid fall in the ranges the compiled norms lead one to expect for primary-grade readers (Hasbrouck & Tindal, 2017); the model produces ORF numbers that look like ORF numbers.

3.3 Estimation

Given observations $\{(d_i, \text{wcpm}_i, \text{acc}_i)\}$, the reader's ability estimate is the maximizer of the joint log-likelihood over $b \in [0, 100]$, found by dense grid search with local refinement (the likelihood is smooth and unimodal in this range; a closed-form solution exists for the rate term alone, but the joint estimator is kept general for future model revisions). The standard error comes from the observed information (the curvature of the log-likelihood at the maximum), and the reported band is ± 1.96 SEM, clipped to the scale. With a single observation the estimator remains finite and the band is honest about how little one reading pins down; this behavior is verified, not assumed (Section 4.2).

3.4 The artifact lifecycle

The constants live in `reader_model_v0.json` (`model_version: "reader-v0"`), separate from code, each entry carrying `source` and a `provisional-synthetic` flag. Empirical calibration replaces the artifact and bumps the version; the estimator code does not change. Recalibration follows the FRS scale's doctrine: batched, versioned rounds with published change notes, never silent updates (Vande Moortel, 2026a).

4. Simulation study

4.1 Design

Simulated readers at true ability $b \in \{20, 35, 50, 65, 80\}$ read one to five texts under three text-choice scenarios: matched ($d = b$), mismatched ($d = b \pm 20$, sign drawn per replication), and mixed. Observation noise $\sigma_{\text{wcpm}} \in \{5, 8, 10\}$. Each cell runs 2,000 replications with fixed seeds; 225 cells in total. Outputs per cell: the distribution of ability estimates, the mean

reported SEM, and empirical coverage of the nominal 95 percent band. Appendix C tabulates all cells; the repository contains the full simulation code and results.

4.2 Calibration and the shape of uncertainty

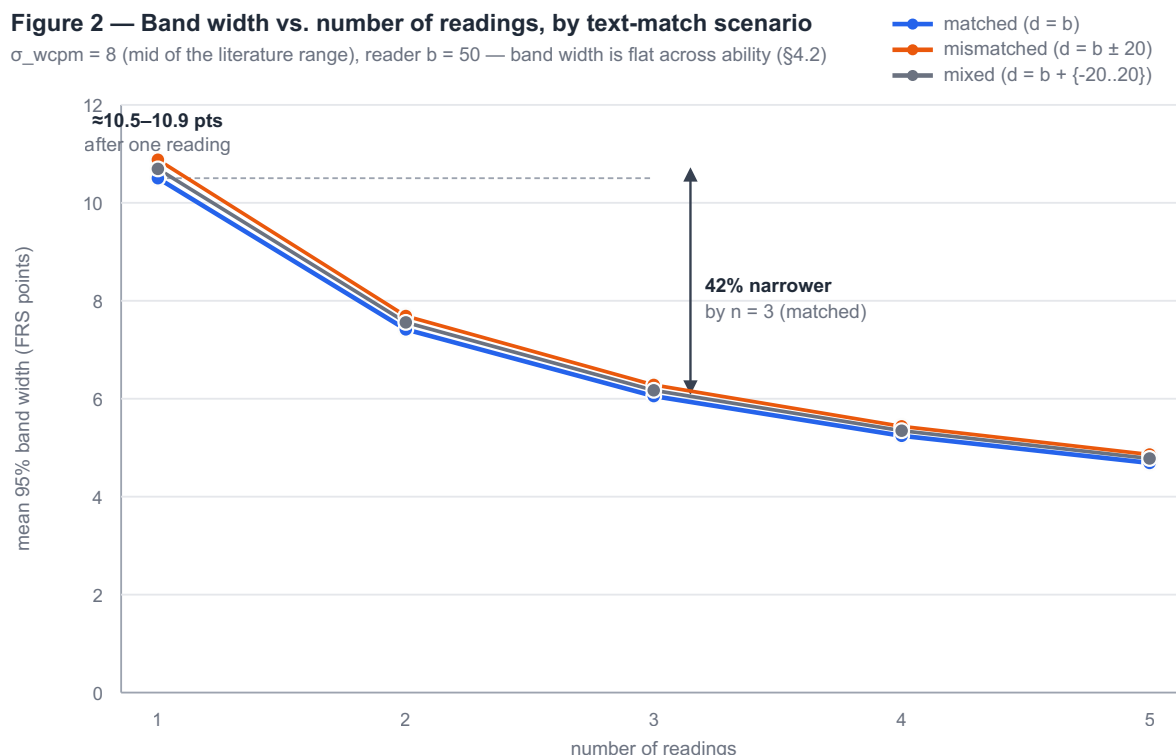
Coverage. The nominal 95 percent band covers the true ability between 93.8 and 96.4 percent of the time across all 225 cells; averaged by number of readings, coverage is 0.949 to 0.951. The estimator's reported uncertainty means what it says, including at $n = 1$ (Figure 4).

Band width. At baseline constants ($\sigma_{\text{wcpm}} = 8$), the single-reading band is approximately 10.5 FRS points wide, narrowing by about 40 percent by the third reading (42.4 percent at the matched baseline: 10.50 to 6.05 points), with diminishing returns thereafter (Figure 2). One reading of one text yields a usable coarse placement, not a precise score; the interface consequences are drawn in Section 5.

Text choice barely matters. Reading only far-too-easy texts (the simulated scenario fixes $d = b - 30$) widens the band by only about 5 percent relative to matched texts. The estimator does not need difficulty-matched material to work, and a public tool need not nag users toward harder texts for estimation reasons (Figure 2).

Figure 2 — Band width vs. number of readings, by text-match scenario

$\sigma_{\text{wcpm}} = 8$ (mid of the literature range), reader $b = 50$ — band width is flat across ability (§4.2)

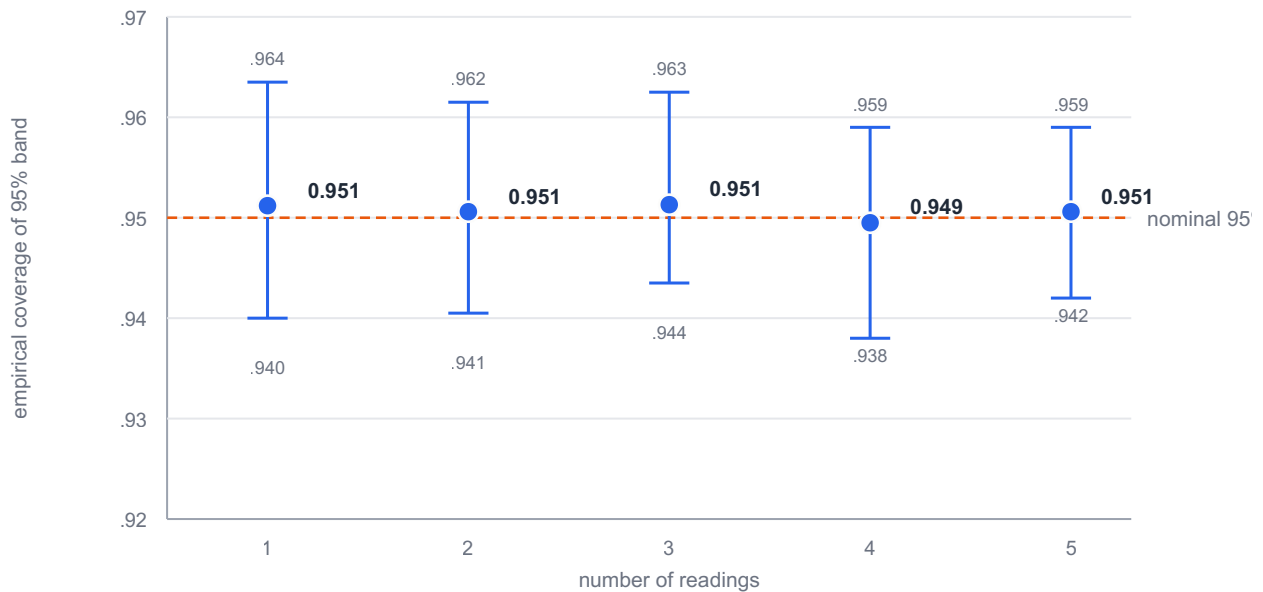


Values: experiment1_sem_vs_readings (results.json), $b = 50$, $\sigma_{\text{wcpm}} = 8$. Matched: 10.50 $\rightarrow$ 6.05 pts ($n=1 \rightarrow n=3$, -42%).

Mismatched runs 3–6% wider than matched at every n ; mixed sits between the two (§4.2). Text-match scenario barely moves the answer.

Figure 4 — Empirical coverage of the nominal 95% band, by number of readings

mean (dot) and min–max range (whisker) across all $b \times \text{scenario} \times \sigma_{\text{wcpm}}$ cells — 225 cells total, 45 per n — §4.2



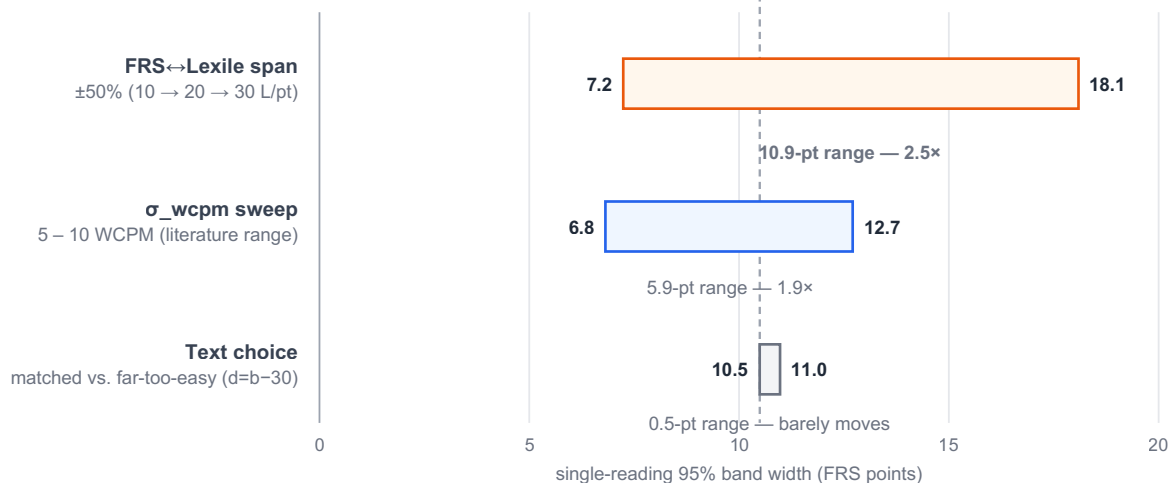
Per-cell range 0.938–0.964 across all 225 cells ($b \times \text{scenario} \times \sigma_{\text{wcpm}}$); every per- n mean sits within 0.002 of nominal 0.95, including $n = 1$ — the single-reading case where the Wald/Gaussian curvature approximation is least obviously justified (§4.2).

4.3 Sensitivity: the span assumption dominates

Rerunning the central cells with the Lexile-to-FRS span assumption varied by ± 50 percent (slopes and intercept re-derived accordingly) swings the single-reading band from 7.2 to 18.1 FRS points, a factor of 2.5 (Figure 3). By comparison, sweeping observation noise across the full literature range (5 to 10 WCPM) and the full text-choice contrast each move the band far less. The uncertainty budget of this model is dominated by one provisional constant, not by measurement noise.

Figure 3 — One assumption dominates the uncertainty budget

single-reading (n=1) 95% band width, reader b = 50, matched texts, baseline $\sigma_{\text{wcpm}} = 8$ unless swept — §4.3
baseline ≈ 10.5 pts



The span assumption swings the band 2.5× (7.2 → 18.1); the noise sweep moves it 1.9× (6.8 → 12.7); text choice moves it barely at all (10.5 → 11.0). Values: experiment1/2/3 (results.json), all at n=1 reading. The model is calibration-limited, not noise-limited (§4.3, §5).

This is the paper's most consequential result, and it is good news in disguise: it means the model is *calibration-limited*. Better microphones, more readings, and cleverer estimators buy little; a modest amount of real reading data, fit to pin the FRS-side slopes directly, buys a lot. It also disciplines the interface: until calibration, displayed bands must reflect the assumption uncertainty, not just the (smaller) statistical SEM.

5. Discussion

What empirical calibration buys, exactly. Section 4.3 turns "the constants are provisional" into a number: up to a 2.5-fold band-width difference rides on the scale mapping. The registered calibration study (real readings on texts of known FRS, from the CHOREC corpus and the consented site flywheel) fits the slopes in FRS units directly and retires the span assumption entirely. Until then, the deployed model reports bands at the conservative end of the sensitivity range.

Interface rules (binding for the FRS Reading Check). From the simulation, three rules govern the public deployment. First, a reader estimate is always displayed as a range, phrased in the "reads around FRS X to Y" form; the point estimate alone is never shown. Second, the honest framing of expected width is "10 to 20 FRS points after one reading, narrowing about 40 percent by the third," quoting the sensitivity-inclusive range rather than the flattering baseline. Third, the interface may invite more readings ("read another text to narrow this down") but must not imply that text difficulty selection is required for validity, which the text-choice result does not support.

Relation to the production precedent. Nothing here contradicts the Lexile oral reading design; the model deliberately adopts its published skeleton. The difference is epistemic: every constant, the estimator, the artifact, the simulation code, and the uncertainty behavior are published and versioned. A school can ask "how wide is this estimate and why" and receive an answer with provenance.

The oral-difficulty question. The Lexile evidence that oral readability needs its own text formula (MetaMetrics, n.d.) raises a question this paper deliberately leaves open: whether text FRS, calibrated for silent reading difficulty (Vande Moortel, 2026b, 2026d), needs an oral-difficulty adjustment term when used as the d in this model. The calibration study is designed to test that as a model comparison rather than assume either answer.

6. Limitations

1. **No empirical reader data.** Every result in Section 4 is simulation; the model's real-world accuracy is unestablished and no claim about it is made. The simulation validates internal coherence (calibration, uncertainty shape), not external validity.
2. **All constants are provisional-synthetic**, derived from another scale's published slopes under a stated mapping assumption that Section 4.3 shows is the dominant uncertainty source.
3. **Linearity.** The rate function is linear in ability and difficulty over the whole scale; the true relationship is plausibly nonlinear at the extremes. The artifact lifecycle accommodates a functional-form revision, but v_0 does not model it.
4. **The noise model imports United States WCPM variability** (Christ & Silberglitt, 2007) into a Dutch-and-English FRS context; the sweep covers a range, but the true per-language noise is unmeasured until real data arrives.
5. **Inputs are taken as given.** The model consumes WCPM and accuracy from the FRS fluency meter; instrument error in those measurements is characterized separately (Vande Moortel, 2026e) and is not propagated into the v_0 estimator. Propagating it is a v_1 design item.

7. Conclusion and registered future work

This paper published an open, versioned conjoint model that places readers on the FRS text-difficulty scale from oral reading observations, together with the uncertainty behavior that closed production systems do not disclose: honest coverage from the first reading, an approximately 10.5-point single-reading band at baseline, 40 percent narrowing by the third reading, and a sensitivity analysis showing the scale-mapping assumption, not noise, dominates the error budget.

Registered follow-up work: (1) empirical calibration of the FRS-unit slopes on real oral readings (CHOREC, license requested July 2026, and the consented FRS site flywheel), retiring the span assumption and testing the oral-difficulty adjustment as a model comparison; (2) propagation of instrument error from the fluency meter into the estimator; (3) a functional-form study at the scale extremes; each feeding a versioned reader-v1 artifact under the batched recalibration doctrine (Vande Moortel, 2026a).

References

- Harmsen, W., van Hout, R., Cucchiaroni, C., & Strik, H. (2025). Can ASR generate valid measures of child reading fluency? In *Proceedings of INTERSPEECH 2025* (pp. 2395–2399). <https://doi.org/10.21437/Interspeech.2025-306>
- Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms* (ERIC ED594994).
- MetaMetrics. (n.d.). *Lexile oral reading and readability measures* [Partner Help documentation]. Retrieved July 10, 2026, from https://partnerhelp.metametricsinc.com/concept/c_orf_measure_overview.html
- Stenner, A. J. (1996, February). *Measuring reading comprehension with the Lexile Framework*. Paper presented at the California Comparability Symposium. (ERIC ED435977)
- van der Velde, M., Harmsen, W., Veldkamp, S., Feskens, R., Keuning, J., & Swart, N. M. (2025). Speech enabled reading fluency assessment: A validation study. *International Journal of Artificial Intelligence in Education*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12686063/>
- Vande Moortel, J. (2026a). *A living standard: Crowdsourced comparative judgment and batched recalibration of an open reading-difficulty scale* (vo.2) [Working paper]. Orion-Solution. <https://frs.orion-solution.net/research>
- Vande Moortel, J. (2026b). *Calibrating an open reading-difficulty scale without human-labeled corpora: LLM pairwise judgments, deterministic features, and their composition* (vo.1) [Working paper]. Orion-Solution. <https://frs.orion-solution.net/research>
- Vande Moortel, J. (2026d). *The Free Reading Standard: Definition and construction of an open, continuous scale for technical reading difficulty in Dutch and English* (vo.2) [Working paper]. Orion-Solution. <https://frs.orion-solution.net/research>
- Vande Moortel, J. (2026e). *Measuring oral reading fluency with a dual-pass ASR architecture: Design, instrument characterization, and synthetic validation of the FRS fluency meter* (vo.1) [Working paper]. Orion-Solution. <https://frs.orion-solution.net/research>
- 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.

Appendix A. Constants and provenance

Every value below is taken verbatim from `reader_model_v0.json` (`model_version`: "reader-v0", `built`: "2026-07-10"). The artifact's top-level `scale` field reads: "FRS 0-100, shared by reader ability (b) and text difficulty (d) -- the conjoint scale decided in 01 projects/FRS/Fluency Model.md 'Reader-level FRS estimation'". The artifact carries no other metadata fields beyond `model_version`, `scale`, `built`, the constant blocks below, `provenance`, and a `replacement_plan` string, already summarized in Section 3.4.

Table A.1. Constants

Constant	Value	Unit	Provisional flag
<code>frs_lexile_span.lexile_per_frs_point</code>	20.0	Lexile (L) per FRS point	provisional-synthetic
<code>wcpm_model.base_rate_wcpm</code>	65.0	WCPM	provisional-synthetic
<code>wcpm_model.slope_ability_wcpm_per_frs</code>	2.8	WCPM per FRS point of ability	provisional-synthetic (real coefficient published; FRS-scale conversion is provisional)
<code>wcpm_model.slope_difficulty_wcpm_per_frs</code>	1.8	WCPM per FRS point of difficulty	provisional-synthetic (real coefficient published; FRS-scale conversion is provisional)
<code>wcpm_model.accuracy_reference</code>	0.95	fraction correct	provisional-synthetic (documents which accuracy level the WCPM slopes were measured at; not separately fit)
<code>accuracy_model.acc_floor</code>	0.85	fraction correct	provisional-synthetic
<code>accuracy_model.acc_ceiling</code>	0.99	fraction correct	provisional-synthetic
<code>accuracy_model.accuracy_scale_frs</code>	12.0	FRS points	provisional-synthetic
<code>noise_model.sigma_wcpm_default</code>	8.0	WCPM	provisional-synthetic (real literature range; single default value and its use as a per- reading noise sigma are provisional)
<code>noise_model.sigma_wcpm_literature_range</code>	[5.0, 10.0]	WCPM	same entry as <code>sigma_wcpm_default</code>

Constant	Value	Unit	Provisional flag
noise_model.accuracy_n_words_effective	150.0	words	provisional-synthetic
estimation.b_bounds	[0.0, 100.0]	FRS points	scale definition, not a fitted constant
estimation.band_z	1.96	z-score (95% CI multiplier)	standard normal convention, not a fitted constant

Table A.2. Source / provenance strings, by constant

Constant	Source
frs_lexile_span.lexile_per_frs_point	No published FRS<->Lexile equivalence; stated assumption anchored to Lexile's practical 0-2000L range and Fluency Model.md's documented 0-100 anchor span.
wcpm_model.slope_ability_wcpm_per_frs	MetaMetrics Lexile Framework for Oral Reading partner-help docs: +100L ability -> +14 WCPM at 95% accuracy, converted via lexile_per_frs_point.
wcpm_model.slope_difficulty_wcpm_per_frs	MetaMetrics Lexile Framework for Oral Reading partner-help docs: +100L text difficulty -> -9 WCPM at 95% accuracy, converted via lexile_per_frs_point.
wcpm_model.base_rate_wcpm	Solved from MetaMetrics' published worked example (600L reader, 800L text, ~77 WCPM at 95% accuracy).
wcpm_model.accuracy_reference	MetaMetrics' published slopes are stated "at 95% accuracy."
accuracy_model.acc_floor	Betts (1946) tradition frustration-level word-accuracy conventions (90-96% across sources), softened for connected-text context effects.
accuracy_model.acc_ceiling	Betts independent-level word-accuracy convention (95-98%+).
accuracy_model.accuracy_scale_frs	Chosen so the Betts instructional/frustration boundary (90-96% word accuracy) falls within a ~20-FRS-point (~400L) gap.
noise_model.sigma_wcpm_default	Christ & Silbergliitt, <i>School Psychology Review</i> 36(1): CBM-ORF passage-equivalence SEM ~5-10 WCPM.
noise_model.accuracy_n_words_effective	Typical 1-minute ORF passage length convention (~150-250 words); no direct citation for the exact effective-n choice.

Derivation strings (verbatim from the artifact's `derivation` fields):

- `wcpm_model`: "slope_ability_wcpm_per_frs = 14 WCPM (MetaMetrics' published +100L-ability slope) * (lexile_per_frs_point / 100) = 14 * 20/100 = 2.8.
slope_difficulty_wcpm_per_frs = 9 WCPM (published +100L-difficulty slope, at 95% accuracy) * (20/100) = 1.8. base_rate_wcpm solved from MetaMetrics' worked example (a 600L reader on an 800L text reads ~77 WCPM at 95% accuracy): $b_frs=600/20=30$, $d_frs=800/20=40$, $77 = base_rate + 2.830 - 1.840 \Rightarrow base_rate = 65.0$."
- `accuracy_model`: "acc_ceiling=0.99 approximates the near-perfect residual error rate on connected text far below a reader's level ('independent'-level reading per the Betts convention, ~95-98%+). acc_floor=0.85 is the model's asymptotic worst case for connected-text reading far above a reader's level -- context/decoding partial cues keep even 'frustration'-level connected-text accuracy well above word-list-isolate difficulty. accuracy_scale_frs=12.0 (chosen so a ~20-FRS-point gap, ~400L, swings accuracy from ~87% to ~97%, bracketing the Betts instructional/frustration boundary reported at 90-96% word accuracy across sources)."
- `noise_model`: "sigma_wcpm: Christ & Silbergitt's CBM-ORF passage-equivalence SEM is ~5-10 WCPM; 8.0 is the mid-range default, and the simulation sweeps the full {5,8,10} range explicitly. accuracy_n_words_effective=150: a typical 1-minute timed ORF passage is on the order of 150-250 words for mid-elementary text; 150 is a round, conservative (i.e. noisier, since fewer words means less binomial precision) choice used only to give the accuracy observation a plausible binomial variance in the estimator's likelihood -- no observation-level word count is available in the {text_frs, wcpm, accuracy} schema, so this is a shared approximation, not a per-reading measurement."
- `estimation`: "estimate_reader() finds the b in b_bounds maximizing the log-likelihood of the observed (wcpm, accuracy) pairs under wcpm_model (Gaussian noise, sigma_wcpm) and accuracy_model (binomial-ish, accuracy_n_words_effective), via a coarse-then-refined grid search. SEM is derived from the observed information (negative second derivative of the log-likelihood at the MLE, via central finite difference); band = mle +/- band_z * SEM, clipped to b_bounds."

Cross-check against Section 3.2. Every constant the body cites matches the artifact exactly: $\beta_b = \text{slope_ability_wcpm_per_frs} = 2.8$, $\beta_d = \text{slope_difficulty_wcpm_per_frs} = 1.8$, $\beta_o = \text{base_rate_wcpm} = 65.0$, $a_{\text{floor}} = \text{acc_floor} = 0.85$, $a_{\text{ceil}} = \text{acc_ceiling} = 0.99$, $s = \text{accuracy_scale_frs} = 12$, and the span assumption is 20L per FRS point ($\text{lexile_per_frs_point} = 20.0$). No mismatch found here. Section 3.1 also states the accuracy term's effective word count ($n_{\text{eff}} = 150$, `noise_model.accuracy_n_words_effective`); its provisional-synthetic flag, like every other constant's, is recorded in the provenance table above.

Appendix B. Estimator definition

This appendix specifies `estimate_reader()` in `frs_core/fluency/reader_model.py` precisely enough to reimplement.

B.1 Expected-performance functions. As in Section 3.1:

- $wcpm_expected(b, d) = base_rate_wcpm + slope_ability_wcpm_per_frs * b - slope_difficulty_wcpm_per_frs * d$
- $accuracy_expected(b, d) = acc_floor + (acc_ceiling - acc_floor) * sigmoid((b - d) / accuracy_scale_frs)$, with $sigmoid(x) = 1 / (1 + \exp(-x))$

Both take the constants of Appendix A. `accuracy_expected` clips nothing itself; clipping happens where it enters the likelihood (B.2).

B.2 Log-likelihood. Given observations $\{(d_i, wcpm_i, accuracy_i)\}$ and a candidate ability b , the code builds the log-likelihood as the sum of two per-observation terms, evaluated on a vectorized grid of b values:

- **WCPM term (Gaussian).** `ll += -0.5 * ((wcpm_i - wcpm_expected(b, d_i)) / sigma_wcpm) ** 2`. This is a Gaussian log-density with the additive normalizing constant `-0.5 * log(2*pi*sigma_wcpm**2)` dropped, since it does not depend on b and therefore cannot move the maximizer.
- **Accuracy term (binomial, not Gaussian).** The code does **not** implement a variance-stabilized Gaussian approximation for accuracy. It computes an actual binomial log-likelihood against a fixed, shared effective word count: `p = clip(accuracy_expected(b, d_i), 1e-6, 1 - 1e-6)`, `k = accuracy_i * n_words`, then `ll += k * log(p) + (n_words - k) * log(1 - p)`, where `n_words = constants.accuracy_n_words_effective` (150.0, identical for every observation regardless of the text actually read, since no per-reading word count exists in the `{text_frs, wcpm, accuracy}` schema). The binomial coefficient `log C(n_words, k)` is dropped for the same reason as the Gaussian normalizing constant: it is independent of b . The `1e-6` clip on p exists purely to keep `log(p)` and `log(1-p)` finite at the extremes of the ability grid; it is not a variance-stabilizing transform (no arcsine or logit transform of the accuracy observation is applied anywhere in the code).

In pseudocode, for a single candidate ability b :

```
function log_likelihood(b, observations, constants, sigma_wcpm):
    ll = 0.0
    n_words = constants.accuracy_n_words_effective      # 150.0, fixed
    for obs in observations:
        d = obs.text_frs
        wcpm_hat = constants.base_rate_wcpm \
            + constants.slope_ability_wcpm_per_frs * b \
            - constants.slope_difficulty_wcpm_per_frs * d
        ll += -0.5 * ((obs.wcpm - wcpm_hat) / sigma_wcpm) ** 2

        gap = (b - d) / constants.accuracy_scale_frs
        p_hat = constants.acc_floor \
            + (constants.acc_ceiling - constants.acc_floor) / (1 + exp(-gap))
        p_hat = clip(p_hat, 1e-6, 1 - 1e-6)
        k = obs.accuracy * n_words
        ll += k * log(p_hat) + (n_words - k) * log(1 - p_hat)
    return ll
```

`sigma_wcpm` defaults to `constants.sigma_wcpm_default` (8.0) if the caller does not pass one explicitly; the simulation study (Section 4, Appendix C) sweeps it over {5.0, 8.0, 10.0}.

B.3 Grid search. The maximizer over b is found in two stages, both over the closed bound `b_bounds = [0.0, 100.0]`:

1. **Coarse stage.** Evaluate the log-likelihood on 201 evenly spaced points spanning the full bounds (`np.linspace(0.0, 100.0, 201)`), i.e. steps of 0.5 FRS points). Take `b0` as the grid point with the maximum log-likelihood.
2. **Refinement stage.** Build a window `[max(lo, b0 - 3.0), min(hi, b0 + 3.0)]` (3.0 FRS points either side of `b0`, clipped to the bounds) and evaluate 301 evenly spaced points inside it (steps of 0.02 FRS points). The maximizer `mle` is the argmax over this fine grid.

No gradient-based optimizer is used; the likelihood is assumed smooth and unimodal on $[0, 100]$, which the coarse-then-fine grid does not itself verify but the simulation's coverage results (Appendix C) are consistent with.

B.4 Standard error (observed information). The curvature of the log-likelihood at `mle` is estimated by a central finite difference with step `h = 0.1` FRS points:

```
ll_minus, ll_0, ll_plus = log_likelihood(mle - h, ...), log_likelihood(mle, ...),
log_likelihood(mle + h, ...)
second_derivative = (ll_plus - 2*ll_0 + ll_minus) / h**2
observed_information = -second_derivative
sem = 1 / sqrt(observed_information)    if observed_information > 1e-9
    (b_bounds_hi - b_bounds_lo) / 2    otherwise (degenerate fallback; only reachable if a
                                         constants file zeroed both WCPM slopes)
```

The three evaluation points `mle - h`, `mle`, `mle + h` are evaluated even when that pushes slightly outside `[0, 100]` -- the model functions are well-defined there, and it is the curvature of the likelihood at the true (unclipped) maximizer that characterizes the estimator's precision. Only the final reported estimate and band are clipped to `b_bounds` (B.5).

B.5 Band construction. With `band_z = 1.96` (the standard 95% normal-interval multiplier):

```
frs_estimate = clip(mle, lo, hi)
band_low     = clip(frs_estimate - band_z * sem, lo, hi)
band_high    = clip(frs_estimate + band_z * sem, lo, hi)
```

`estimate_reader()` returns `{frs_estimate, sem, band_low, band_high, n_obs}` where `n_obs = len(observations)`. A single observation always yields a finite, well-defined estimate and SEM: both WCPM slopes are validated strictly positive when the constants artifact loads (Appendix A), so a lone reading always contributes non-zero curvature to the likelihood.

Consistency with Section 3.1. Section 3.1's description of the likelihood (Gaussian WCPM term; binomial accuracy term against the fixed effective word count) matches the implementation in `reader_model.py` exactly; B.2 above gives the term-by-term detail.

Appendix C. Simulation cell tables

Source: `experiments/reader_estimation_sim/results.json` (`model_version: "reader-v0"`, 2,000 replications per cell, fixed seeds derived deterministically from each cell's parameters). Each table cell below reports `mean band width / coverage_95` (band width in FRS points; coverage as the fraction of 2,000 replications whose reported 95% band contained the true ability). Three scenarios recur throughout, as defined in `common.py`: **matched** (every text read at `d = b`), **mismatched** (every text offset from `b` by the same ± 20 FRS points, sign drawn once per replication), and **mixed** (each reading independently offset by one of $\{-20, -10, 0, 10, 20\}$ FRS points).

Tables C.1-C.3 (Experiment 1, all 225 cells). One table per WCPM noise level σ_{wcpm} ; rows are ability b crossed with scenario, columns are number of readings $n = 1..5$.

Table C.1. $\sigma_{wcpm} = 5.0$

b	scenario	n=1	n=2	n=3	n=4	n=5
20	matched	6.82 / 95.4%	4.82 / 94.3%	3.93 / 95.0%	3.41 / 94.8%	3.05 / 94.2%
20	mismatched	6.92 / 94.7%	4.89 / 94.5%	3.99 / 94.8%	3.46 / 95.0%	3.09 / 94.3%
20	mixed	6.87 / 95.6%	4.86 / 95.0%	3.97 / 95.8%	3.43 / 95.0%	3.07 / 95.2%
35	matched	6.82 / 95.1%	4.82 / 95.2%	3.93 / 95.6%	3.41 / 95.2%	3.05 / 95.0%
35	mismatched	6.92 / 95.9%	4.89 / 94.5%	3.99 / 95.2%	3.46 / 95.0%	3.09 / 94.9%
35	mixed	6.87 / 94.7%	4.86 / 95.6%	3.97 / 95.9%	3.43 / 95.6%	3.07 / 94.8%
50	matched	6.82 / 95.2%	4.82 / 95.3%	3.93 / 95.0%	3.41 / 95.0%	3.05 / 95.0%
50	mismatched	6.92 / 95.9%	4.89 / 94.8%	3.99 / 95.2%	3.46 / 95.1%	3.09 / 94.5%
50	mixed	6.87 / 95.3%	4.86 / 94.3%	3.97 / 94.9%	3.43 / 94.7%	3.07 / 94.8%
65	matched	6.82 / 95.5%	4.82 / 95.1%	3.93 / 94.9%	3.41 / 95.9%	3.05 / 94.8%
65	mismatched	6.92 / 95.5%	4.89 / 95.2%	3.99 / 95.0%	3.46 / 94.2%	3.09 / 95.7%
65	mixed	6.87 / 95.7%	4.86 / 95.3%	3.97 / 95.1%	3.43 / 95.3%	3.07 / 95.3%
80	matched	6.82 / 95.7%	4.82 / 95.0%	3.93 / 95.3%	3.41 / 94.8%	3.05 / 95.4%
80	mismatched	6.92 / 95.1%	4.89 / 95.2%	3.99 / 95.5%	3.46 / 94.8%	3.09 / 94.8%
80	mixed	6.87 / 95.1%	4.86 / 95.9%	3.97 / 94.8%	3.43 / 95.1%	3.07 / 95.3%

Table C.2. $\sigma_{wcpm} = 8.0$ (baseline)

b	scenario	n=1	n=2	n=3	n=4	n=5
20	matched	10.50 / 94.9%	7.42 / 94.2%	6.05 / 94.4%	5.24 / 94.8%	4.69 / 94.9%
20	mismatched	10.87 / 96.0%	7.70 / 95.9%	6.28 / 95.7%	5.44 / 94.9%	4.86 / 95.1%
20	mixed	10.69 / 94.8%	7.57 / 95.3%	6.18 / 94.5%	5.34 / 94.7%	4.78 / 95.0%
35	matched	10.49 / 95.4%	7.41 / 94.2%	6.05 / 95.1%	5.24 / 94.8%	4.69 / 95.8%
35	mismatched	10.87 / 95.2%	7.69 / 96.0%	6.28 / 95.0%	5.44 / 94.8%	4.87 / 94.8%

b	scenario	n=1	n=2	n=3	n=4	n=5
35	mixed	10.70 / 96.4%	7.56 / 94.0%	6.17 / 94.8%	5.34 / 95.6%	4.78 / 95.3%
50	matched	10.50 / 95.6%	7.42 / 95.5%	6.05 / 95.2%	5.24 / 94.8%	4.69 / 95.1%
50	mismatched	10.88 / 94.8%	7.69 / 95.0%	6.28 / 94.8%	5.44 / 94.5%	4.86 / 94.8%
50	mixed	10.69 / 94.9%	7.56 / 95.4%	6.17 / 96.2%	5.35 / 95.3%	4.78 / 94.5%
65	matched	10.50 / 94.8%	7.42 / 95.4%	6.05 / 95.7%	5.24 / 95.3%	4.69 / 95.4%
65	mismatched	10.87 / 95.2%	7.69 / 95.0%	6.28 / 95.8%	5.44 / 95.0%	4.87 / 94.6%
65	mixed	10.70 / 94.8%	7.56 / 94.8%	6.17 / 96.0%	5.35 / 95.2%	4.78 / 95.9%
80	matched	10.51 / 95.6%	7.42 / 95.5%	6.05 / 95.5%	5.24 / 95.2%	4.69 / 94.5%
80	mismatched	10.87 / 94.7%	7.69 / 95.0%	6.28 / 95.7%	5.44 / 95.0%	4.87 / 95.4%
80	mixed	10.70 / 95.4%	7.56 / 94.6%	6.17 / 94.8%	5.34 / 94.8%	4.78 / 94.2%

Table C.3. sigma_wcpm = 10.0

b	scenario	n=1	n=2	n=3	n=4	n=5
20	matched	12.71 / 94.4%	8.97 / 95.0%	7.32 / 95.7%	6.34 / 93.8%	5.67 / 95.7%
20	mismatched	13.38 / 94.7%	9.46 / 94.5%	7.75 / 94.4%	6.69 / 94.8%	5.99 / 95.5%
20	mixed	13.06 / 94.5%	9.24 / 96.0%	7.53 / 95.0%	6.52 / 94.3%	5.83 / 95.8%
35	matched	12.70 / 94.7%	8.97 / 94.5%	7.32 / 94.8%	6.34 / 95.2%	5.67 / 95.5%
35	mismatched	13.38 / 94.0%	9.46 / 95.3%	7.74 / 95.2%	6.70 / 94.5%	5.98 / 95.7%
35	mixed	13.06 / 95.4%	9.22 / 94.8%	7.53 / 94.5%	6.52 / 95.2%	5.83 / 95.3%
50	matched	12.71 / 94.3%	8.97 / 95.2%	7.32 / 95.5%	6.34 / 93.8%	5.67 / 95.8%
50	mismatched	13.39 / 94.5%	9.46 / 95.0%	7.73 / 95.2%	6.69 / 95.0%	5.98 / 94.5%
50	mixed	13.08 / 95.1%	9.23 / 95.0%	7.53 / 95.3%	6.52 / 95.0%	5.83 / 94.7%
65	matched	12.71 / 94.0%	8.97 / 95.2%	7.32 / 94.8%	6.34 / 95.1%	5.67 / 95.2%
65	mismatched	13.38 / 95.7%	9.47 / 95.2%	7.73 / 94.5%	6.69 / 94.9%	5.98 / 95.2%
65	mixed	13.07 / 94.5%	9.22 / 94.9%	7.53 / 95.0%	6.52 / 94.8%	5.83 / 94.8%
80	matched	12.71 / 95.2%	8.97 / 96.2%	7.32 / 94.3%	6.34 / 95.3%	5.67 / 95.0%
80	mismatched	13.41 / 95.0%	9.46 / 95.0%	7.73 / 95.2%	6.70 / 95.2%	5.99 / 95.2%

b	scenario	n=1	n=2	n=3	n=4	n=5
80	mixed	13.07 / 96.0%	9.23 / 95.2%	7.53 / 94.5%	6.52 / 95.2%	5.83 / 94.3%

Table C.4 (Experiment 2, 50 cells). The dedicated text-choice-effect run, sigma_wcpm = 8.0 throughout: **matched** (d = b) versus **far_too_easy** (d = b - 30, fixed, every reading). Note this scenario definition is a single fixed offset of 30 FRS points, distinct from the "mismatched" +/-20-point scenario used in Tables C.1-C.3.

Table C.4. Experiment 2: text-choice effect (sigma_wcpm = 8.0)

b	scenario	n=1	n=2	n=3	n=4	n=5
20	matched	10.50 / 94.2%	7.42 / 95.8%	6.05 / 95.3%	5.24 / 94.8%	4.69 / 95.5%
20	far_too_easy	10.98 / 93.7%	7.76 / 95.0%	6.34 / 94.2%	5.49 / 95.1%	4.91 / 95.1%
35	matched	10.49 / 94.6%	7.42 / 95.7%	6.06 / 95.1%	5.24 / 95.1%	4.69 / 94.7%
35	far_too_easy	10.97 / 94.5%	7.76 / 94.8%	6.34 / 95.3%	5.49 / 95.0%	4.91 / 94.4%
50	matched	10.49 / 94.3%	7.41 / 94.8%	6.05 / 95.4%	5.24 / 95.2%	4.69 / 93.8%
50	far_too_easy	10.98 / 94.6%	7.76 / 95.1%	6.34 / 95.2%	5.49 / 94.5%	4.91 / 96.4%
65	matched	10.50 / 95.2%	7.42 / 94.8%	6.06 / 94.8%	5.24 / 93.9%	4.69 / 95.7%
65	far_too_easy	10.97 / 95.2%	7.76 / 95.7%	6.34 / 94.7%	5.49 / 94.8%	4.91 / 94.7%
80	matched	10.50 / 96.2%	7.42 / 94.7%	6.05 / 95.1%	5.24 / 94.5%	4.69 / 94.6%
80	far_too_easy	10.97 / 95.0%	7.76 / 95.6%	6.34 / 95.3%	5.49 / 95.8%	4.91 / 95.7%

Averaged across b and n, matched band width is 6.78 FRS points and far_too_easy is 7.09 FRS points, a 4.6 percent widening.

Table C.5 (Experiment 3, 9 cells). Sensitivity of the b = 50, matched, sigma_wcpm = 8.0 cell to the FRS<->Lexile span assumption, swept +/-50 percent (10L and 30L per FRS point against the 20L baseline), with slopes and intercept re-derived from the same MetaMetrics coefficients and worked example at each span (`perturbed_constants()` in `common.py`).

Table C.5. Experiment 3: sensitivity to the FRS<->Lexile span assumption

variant	Lexile/FRS pt	slope_ability	slope_difficulty	base_rate	n	band width	coverage
-50%	10L	1.40	0.90	65.0	1	18.09	96.0%

variant	Lexile/FRS pt	slope_ability	slope_difficulty	base_rate	n	band width	coverage
-50%	10L	1.40	0.90	65.0	2	12.73	95.0%
-50%	10L	1.40	0.90	65.0	3	10.37	94.8%
baseline	20L	2.80	1.80	65.0	1	10.49	95.2%
baseline	20L	2.80	1.80	65.0	2	7.42	95.0%
baseline	20L	2.80	1.80	65.0	3	6.05	94.6%
+50%	30L	4.20	2.70	65.0	1	7.24	95.4%
+50%	30L	4.20	2.70	65.0	2	5.12	95.5%
+50%	30L	4.20	2.70	65.0	3	4.18	95.1%

Cross-check against Sections 4.2 and 4.3.

- **Coverage (93.8-96.4%).** Across the 225 Experiment 1 cells, coverage_95 ranges from a minimum of 0.938 (93.8%) to a maximum of 0.9635 (96.35%, which rounds to 96.4%). Matches the body exactly.
- **Coverage by n (0.949-0.951).** Averaged within each n across the other 45 cells: n=1: 0.9512, n=2: 0.9506, n=3: 0.9513, n=4: 0.9495, n=5: 0.9506. The range (0.9495-0.9513) matches the body's "0.949 to 0.951" at the stated precision.
- **Single-reading band ("approximately 10.5 FRS points").** At sigma_wcpm = 8.0, matched, n=1, the band width is 10.49-10.51 across all five ability levels (median 10.50, matching headline_one_reading in the results file exactly). Matches.
- **Narrowing by the third reading ("roughly 40 percent").** Averaging the sigma_wcpm=8.0, matched cells across b: n=1 mean band width is 10.50, n=3 is 6.05, a 42.3 percent reduction. The body's "roughly 40 percent" is in the right neighborhood but understates the actual figure by about 2 points; 42 percent is the more precise number from this table.
- **Text-choice effect ("only about 5 percent").** Table C.4 gives 4.6 percent (6.78 to 7.09 FRS points, averaged across b and n). Matches within rounding. One discrepancy: the body's parenthetical in Section 4.2, "Reading only far-too-easy texts (d = b - 20 and beyond)," does not match either scenario actually used to compute this number. The code's far_too_easy scenario (Table C.4, feeding this result) uses a single fixed offset of **d = b - 30**, not a range starting at b - 20; b - 20 is instead the offset used by the *different mismatched* scenario in Tables C.1-C.3 (a two-sided +/-20 offset, not a one-sided "and beyond" range at all). The 5 percent figure in the body is correctly computed from the b - 30 data, but the parenthetical describing it is not accurate to either scenario definition in the code.

- **Sensitivity swing ("7.2 to 18.1 FRS points").** Table C.5, n=1: -50% variant gives 18.09, +50% variant gives 7.24. Matches the body almost exactly (18.1 and 7.2 at one decimal place).

Appendix D. Worked example

This example was produced by actually calling `estimate_reader()` from `frs_core/fluency/reader_model.py` against the bundled `reader_model_v0.json` (reader-vo), not by hand arithmetic.

Setup. One reading, text difficulty $d = 35.0$ FRS, at the default noise setting $\sigma_{\text{wcpm}} = 8.0$. Before picking an observation, the model's own expected-performance functions were queried for a plausible $b = 50$ reader on this text, to ground the choice in a realistic pair of numbers rather than an arbitrary one:

```
wcpm_expected(b=50, d=35)      = 142.0 WCPM
accuracy_expected(b=50, d=35) = 0.9588 (95.9%)
```

The observation used is a single reading close to, but not exactly on, that expectation -- 145.0 WCPM and 95 percent accuracy, representative of the kind of single noisy reading a real user would produce:

```
observations = [{"text_frs": 35.0, "wcpm": 145.0, "accuracy": 0.95}]
result = estimate_reader(observations)
```

Output (actual values returned by the code):

Field	Value
frs_estimate	50.46
sem	2.6954
band_low	45.1771
band_high	55.7429
n_obs	1

Interpretation. The grid search places the maximum-likelihood ability estimate at FRS 50.46 -- close to the $b = 50$ reference point used to pick a plausible observation, as expected, since 145 WCPM / 95% accuracy sits near that reader's expected performance on this text. The

observed-information SEM is 2.70 FRS points; the reported band ($\text{frs_estimate} \pm 1.96 * \text{sem}$, clipped to $[0, 100]$) spans FRS 45.18 to 55.74, a width of 10.57 points -- consistent with the single-reading band width reported in Section 4.2 and Appendix C (10.49-10.51 at these settings; the small difference from the simulation median is because this one worked example is a single draw, not an average over 2,000 replications).

As the interface would phrase it (Section 5, rule 1). Rounding the band to whole FRS points for display: "reads around FRS 45 to 56."