

A Living Standard: Crowdsourced Comparative Judgment and Batched Recalibration of an Open Reading-Difficulty Scale

Jasper Vande Moortel · Orion-Solution · Version 0.2 (draft) · July 2026

Abstract

An open measurement standard faces a maintenance dilemma that proprietary scales solve with money: a frozen scale decays as language, text production, and coverage needs change, while a continuously updated scale destroys the comparability that makes a standard worth having. This paper specifies how the Free Reading Standard (FRS) resolves the dilemma: a public contribution intake designed around comparative-judgment research, feeding batched, versioned recalibration rounds that leave the scale frozen between releases. The intake never asks laypeople for scale values. Everyone answers one anchored audience question; typed AVI or CEFR labels are offered only to contributors who claim them; and the highest-value signal is a single relative judgment, one A/B comparison between the contributed text and a calibration anchor, entering the same Bradley-Terry pool that calibrated the scale. Design parameters come from the comparative-judgment literature: reliability dosage from meta-analysis (10–14 comparisons per item for $r \approx .70$), non-adaptive pair scheduling to avoid documented reliability inflation, and per-rater quality modeling in the Crowd-BT tradition. A simulation study (1,080 seeded runs, vote noise fitted from the scale's real judgment records) quantifies the mechanism before deployment: median drift on the released anchors stays at or below 0.015 FRS points per three rounds even under 20% adversarial contributions with poor label quality; the quarantine gate intercepts about half of adversarial input, and safety derives from anchor pinning and batching rather than from perfect detection; and placement accuracy improves only when label quality is decent, not from volume alone, which is the empirical justification for the intake design. The intake specified here is deployed at the standard's public site, replacing a simpler interim form that predated this design; the first operational round will test this specification.

Keywords: comparative judgment, crowdsourcing, Bradley-Terry, scale maintenance, readability, open standards

1. Introduction

The Free Reading Standard places Dutch and English texts on an open, continuous 0–100 scale of technical reading difficulty, specified in a companion paper (Vande Moortel, 2026a)

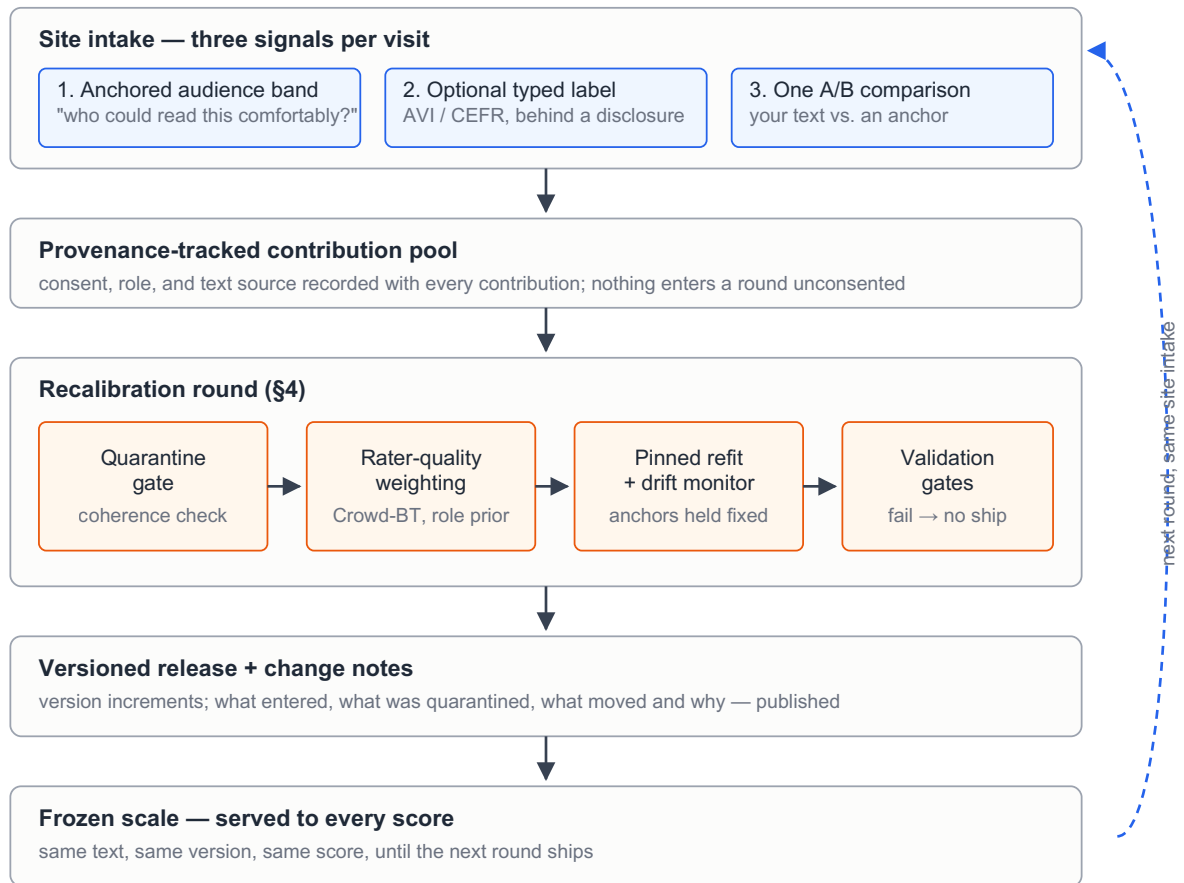
and validated in a second (Vande Moortel, 2026b). Its calibration rests on anchor texts and pairwise difficulty judgments. That construction answers how an open scale gets built. This paper answers a harder question: how does it stay alive?

The dilemma is structural. A scale frozen at release decays: the anchor sets have coverage gaps (the Dutch calibration, for instance, currently has no anchors above the upper-academic region), text production changes, and anchors themselves can drift in meaning, a phenomenon the equating literature names item parameter drift (Kolen & Brennan, 2014). But the naive cure is worse than the disease. A scale that silently updates whenever new data arrives destroys exactly the property users rely on: that a score of 34 means the same thing in June as it did in January. Proprietary scales resolve this with continuous psychometric maintenance programs funded by license fees. An open, free standard needs a mechanism that is itself open, cheap, and safe against the failure modes of public contribution, including spam and adversarial input.

FRS's answer has three parts, and this paper specifies all of them before the mechanism has collected real data, deliberately: publishing the design first makes the first operational round a test of a stated specification rather than a post-hoc rationalization.

1. **An intake design** grounded in comparative-judgment research: contributors are never asked for scale values; they answer an anchored audience question, optionally supply typed expert labels, and optionally give one relative judgment against a served anchor excerpt.
2. **A batched, versioned recalibration algorithm:** contributions accumulate in a provenance-tracked pool; periodically a round runs (quarantine screening, rater-quality weighting, a blended refit with the original anchors pinned, validation gates) and ships as a new scale version with published change notes. Between rounds the scale is frozen. Consistency is a guarantee, not an aspiration.
3. **A simulation study** quantifying the mechanism's safety envelope and benefit conditions across contribution volumes, label-quality levels, and adversarial fractions, with vote noise calibrated from the scale's real judgment records.

Figure 1 gives the mechanism whole, before the sections below take it apart: an intake collects three signals, contributions sit in a provenance-tracked pool, a round processes them behind four gates, a versioned release ships with change notes, and the scale stays frozen until the next round starts the loop again.



Nothing changes between rounds: the frozen scale in Box E is what every analyzer call reads. A round only runs on a batched, versioned schedule — never continuously — which is what keeps the standard comparable.

Figure 1. The living-standard cycle: site intake feeds a provenance-tracked pool, a recalibration round processes it behind four gates (\$4), and a versioned release freezes the scale until the next round.

2. Background

2.1 Comparative judgment: from Thurstone to assessment practice

The method's foundation is Thurstone's (1927) law of comparative judgment: reliable interval scales can be constructed from many binary "which is more X" decisions, because humans are far better at comparing two stimuli than at rating one absolutely. Pollitt (2012) brought the paradigm into educational assessment as Adaptive Comparative Judgement, replacing analytic marking of complex work with holistic paired comparisons. For readability specifically, the CLEAR corpus demonstrated the paradigm at scale, deriving continuous difficulty scores for 4,724 texts from 111,347 pairwise teacher judgments (Crossley et al., 2023), and FRS's own calibration adapted that construction with a language-model judge (Vande Moortel, 2026b).

Two findings from this literature set concrete parameters for any contribution system. First, reliability dosage: a meta-analysis across 49 comparative-judgment assessments found that roughly 10 to 14 comparisons per item are needed for scale-separation reliability near .70, and

26 to 37 for .90, with novice judges needing the higher end of each range (Verhavert, Bouwer, Donche, & De Maeyer, 2019). FRS therefore treats a contributed text's comparative signal as calibration-grade only after it has accumulated at least ten votes; below that threshold it contributes coverage, not placement authority. Second, a warning: adaptive pair selection, the "A" in ACJ, inflates the standard reliability statistic. Bramley's analyses reported scale-separation reliability of 0.97 under adaptive scheduling against 0.82 for the same materials under non-adaptive designs, figures that led commercial comparative-judgment platforms to reduce or abandon adaptivity (Bramley, 2015; Bramley & Vitello, 2019; the 2015 figures are cited here through secondary reports, as noted in the bibliography). FRS v1 therefore serves comparison excerpts non-adaptively, by simple difficulty proximity with randomized presentation order, and commits to reporting any future adaptive design's reliability on non-adaptive holdouts.

2.2 The Flemish comparative-judgment tradition

Comparative judgment is not an exotic import for FRS's home market. The D-PAC project at the University of Antwerp developed comparative judgment into a working assessment methodology for writing and professional competences, producing both the reliability meta-analysis cited above and a commercial spin-off (Comproved). A practically important strand of that work concerns judge expertise: comparative judgment tolerates non-expert judges, who reach comparable reliability to experts at the cost of more comparisons per item (reported across the D-PAC literature; the primary source for the strongest version of this claim resisted retrieval and the claim is used here only in its weak form). For a public intake this is the load-bearing scientific permission: teacher votes and lay votes can share one pool, provided the model weights them rather than trusting them equally.

2.3 Crowdsourced judgment quality

Weighting raters is a solved class of problem. Dawid and Skene (1979) introduced joint estimation of true labels and per-rater error characteristics for categorical judgments; Chen, Bennett, Collins-Thompson, and Horvitz (2013) brought the same idea to pairwise comparisons as Crowd-BT, extending Bradley-Terry with a per-rater quality parameter that natively down-weights spammers and adversarial raters. FRS adopts the Crowd-BT formulation for its vote pool, with one design refinement: the contributor's declared role (teacher, parent, author, other) enters as a prior on the quality parameter rather than as a fixed weight, so that the data can promote a careful parent above a careless teacher. To our knowledge, no published system has pooled a language-model judge and human raters inside one rater-quality model of this kind; FRS's rounds do exactly that, and the claim is stated as an implementation novelty pending one further literature check rather than as an established first.

2.4 Scale maintenance in operational testing

Operational testing programs maintain scales across forms and years through equating and linking designs built on anchor items (Kolen & Brennan, 2014). Two lessons transfer. Anchors are the spine of comparability: FRS rounds never refit the released anchors freely; they are pinned, and the round monitors them for drift instead of letting them move silently. And drift is expected, not exceptional: item parameter drift is a named, monitored phenomenon in that literature, so the FRS round algorithm treats "an anchor no longer sits where it did" as a detectable event that triggers review, not as noise to absorb. The versioned-release discipline itself (frozen between rounds, changelogs on release) is standard practice in measurement programs and in engineering; its application to a public, crowd-fed readability scale is, as far as we could establish, without direct educational precedent, with the closest analogues appearing in continuously refreshed AI evaluation benchmarks, cited here as analogy only.

2.5 What FRS already is

For self-containment: FRS scores texts through a deterministic linguistic feature extractor and a small regression model per language, calibrated against license-clean anchor texts placed by pairwise judgments aggregated with Bradley-Terry, mapped to an anchored 0–100 scale, and served with uncertainty bands and version identifiers (Vande Moortel, 2026a, 2026b). Serving is deterministic and model-free; all parameters, prompts, anchors, and artifacts are published. The standard's public site offers a free analyzer; the contribution intake specified next is designed to sit directly beneath its results. A disclosure for the record: the site initially shipped with an interim form that predated this design and asked contributors for a numeric level directly, the very practice §3's first principle rejects; it was replaced by the design below during the drafting of this paper, and the first recalibration round will accept only v2-format contributions (interim-format submissions are retained but enter rounds as coverage-only texts, never as labels).

3. The intake design

The design question is what to ask a stranger who has just analyzed a text, such that the answer is worth feeding into a measurement standard. Three principles, each traceable to evidence, shaped the answer.

Never ask laypeople for scale values. A visitor who met the FRS scale ninety seconds ago cannot place a text on it; a request for "the AVI level" is jargon to everyone except Dutch primary-school professionals; and absolute ratings are noisy even from experts, which is the founding observation of the comparative-judgment tradition (§2.1) and was reproduced in FRS's own calibration experiments, where absolute scoring by a strong judge compressed the scale while relative judgment preserved ordering (Vande Moortel, 2026b).

Anchor every question in the respondent's world. The one question everyone answers is an audience question: *who could read this text comfortably?* Its options are concrete bands (early primary, roughly ages 6–8; mid primary; upper primary; secondary; general adult; specialist/academic; and an explicit "I don't know"). This is how teachers and parents actually think about difficulty, and coarse bands answered reliably beat fine scales answered randomly.

Let experts be experts, voluntarily. Behind a disclosure control labeled "I know this text's level" sit typed label fields: an AVI level dropdown for Dutch texts, a CEFR dropdown, and a free-text field for the label's provenance (a book cover, a method, the contributor's own assessment). Contributors who never open the control never see the jargon; contributors who do supply exactly the kind of typed, attributable label the anchor pipeline can screen.

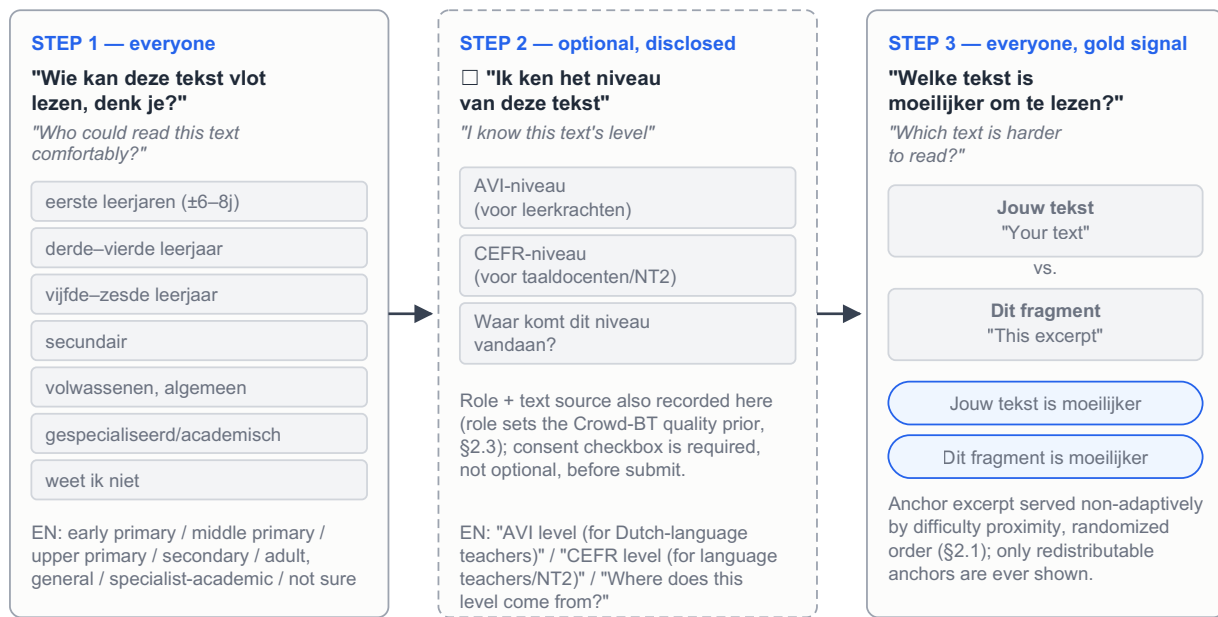
The gold signal is one comparison. After analysis, the site shows one short excerpt from a license-clean calibration anchor near the contributed text's estimated difficulty and asks which of the two is harder to read. That single tap is a human pairwise judgment entering the same Bradley-Terry pool the scale was calibrated with: the visitor becomes an anonymous rater in a Thurstonian design without being asked to understand it. Excerpts are served non-adaptively (§2.1) with randomized presentation order; only redistributable anchors are ever shown.

Every contribution also records the contributor's role and the text's source (own writing, textbook, website, other), the former as a rater-quality prior (§2.3), the latter for copyright routing: only self-authored contributions are candidates for redistribution in future anchor sets; everything else remains internal calibration data. Consent is explicit and required; an optional email enables erasure requests; the privacy architecture (salted IP hashing, rate limiting, retention policy) ships in the standard's published legal documentation. The full form copy in both site languages appears in Appendix A.

Figure 4 shows the deployed flow end to end, in the site's own Dutch copy with an English gloss: the audience-band question every visitor answers, the optional expert-label step behind its disclosure control, and the comparison vote that supplies the gold signal.

The intake flow: everyone answers step 1, experts opt into step 2, everyone sees step 3

Verbatim form copy, Dutch primary with English gloss (site i18n, analyzer.contribute keys)



Source of copy: E:\code\frs\site\src\i18n\locales\{nl,en}.json, keys under analyzer.contribute (Appendix A quotes both languages in full).

Figure 4. The three-step intake flow as deployed: an anchored audience question everyone answers, a disclosed expert-label step, and the pairwise comparison that feeds the Bradley-Terry pool. Verbatim copy in both languages is in Appendix A.

4. The recalibration round algorithm

A round transforms (current version, contribution pool) into (new version, change notes), and nothing changes between rounds. The algorithm, stated compactly (pseudocode in Appendix B):

- 1. Eligibility.** Contributions enter the round if consented and non-quarantined at intake; a text's comparative signal counts only with ≥ 10 accumulated votes (§2.1); typed labels count only with role and provenance recorded.
- 2. Quarantine gate.** Each candidate's signals are checked for internal coherence (band vs. typed label vs. vote-implied position vs. the model's feature-based estimate). Systematic disagreement quarantines the contribution for review rather than deleting it; the gate's purpose is to keep incoherent input out of the fit, not to adjudicate truth.
- 3. Rater-quality weighting.** Votes enter a Crowd-BT fit with per-contributor quality parameters, role as prior. The language-model judge participates as one more rater with its own quality parameter, estimated on the same footing.
- 4. Blended refit with pinned anchors.** The Bradley-Terry fit re-runs over the union of existing judgments and new weighted votes with the released anchors' positions pinned; surviving contributions receive scale positions; the feature-to-scale regression refits per language on the updated target set.

5. **Drift monitor.** The pinned anchors are re-estimated in a shadow fit with pins released; anchors whose shadow position moves beyond a published tolerance are flagged for human review (§2.4). Flagged anchors can be retired or re-labeled in the next round, never silently moved.
6. **Gates and release.** The full validation gate suite from the calibration pipeline re-runs; a round that fails gates does not ship. A shipping round increments the version, publishes change notes (what entered, what was quarantined, what moved and why), and freezes.

The consistency guarantee follows from the structure: scores carry version identifiers; a version's scores never change; and cross-version comparability is maintained through the pinned anchors, exactly the anchor-item logic of operational equating (Kolen & Brennan, 2014).

5. Simulation study

A mechanism that touches a released standard should be stress-tested before it runs on real input. The simulation asks three pre-registered questions: does the round algorithm protect the released scale under hostile input (safety); how much adversarial input does the quarantine gate catch (detection); and under what contribution conditions does the flywheel actually improve the scale (benefit)?

5.1 Design

The full grid crosses 2 languages $\times$ 3 volumes per round (50, 200, 1,000) $\times$ 3 band-label quality levels (correct-band probability 0.9, 0.7, 0.5) $\times$ 2 expert fractions (0.05, 0.2) $\times$ 3 adversarial fractions (0, 0.05, 0.2) $\times$ 10 seeds: 1,080 simulated runs, each executing three cumulative recalibration rounds. Ground truth comes from the real FRS 1.0.1 anchors; synthetic contributed texts interpolate between real anchors in feature space so each carries a known true position. Vote noise is not assumed: the probability of a correct pairwise vote as a function of true difficulty gap was fitted as a logistic from the scale's real judgment records. Adversarial contributors submit inverted or random labels and votes. Two full-grid executions produced byte-identical results; code and outputs are published with the reference implementation (Appendix C).

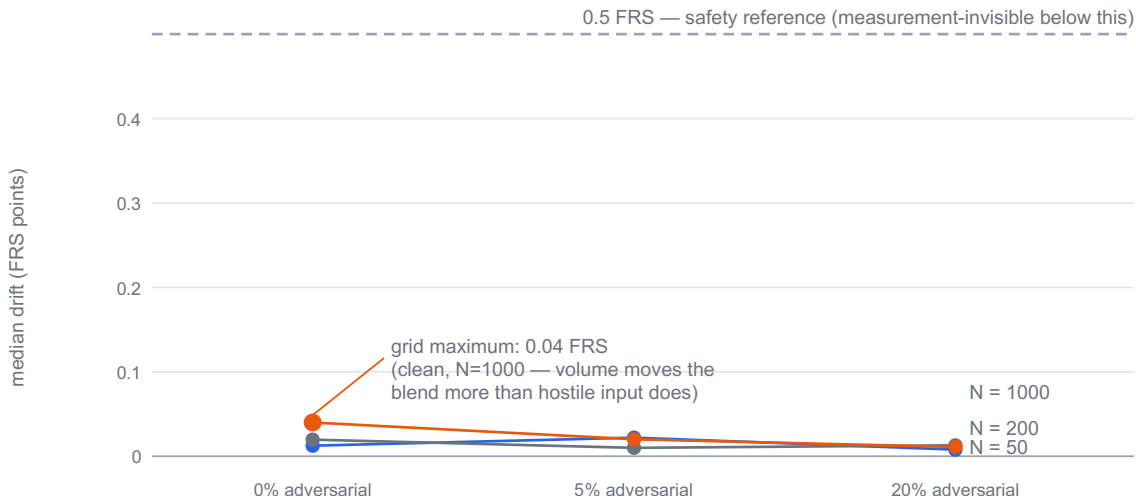
5.2 Results

Safety. Median drift on the released anchors never exceeded 0.04 FRS points in any condition after three rounds, and stayed at or below 0.015 in the joint adversarial worst case (20% adversarial input with 50% band-label accuracy); the grid maximum occurred, counterintuitively, in a clean high-volume condition, where sheer contribution mass moves the blend more

than hostile input does. On a 0–100 scale both figures are measurement-invisible. Safety derives from the pins and the batch structure, not from detection quality.

Median anchor drift stays far below the safety reference

round 3, EN, $p_{\text{band}} = 0.7$, $f_{\text{exp}} = 0.2$ — every grid point, all three volumes

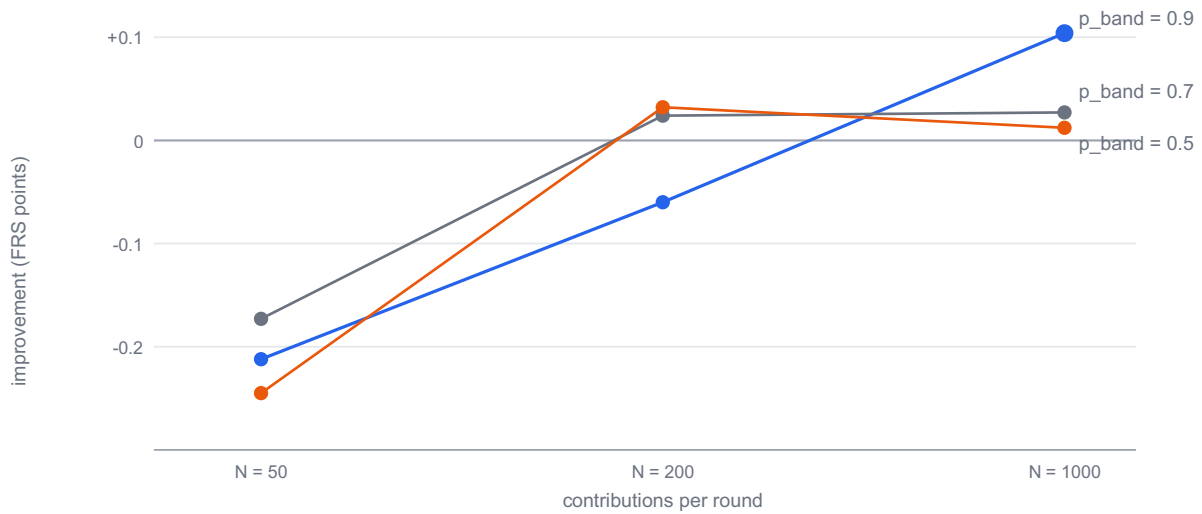


All 9 points (3 volumes $\times$ 3 adversarial fractions) sit under 0.04 FRS points of drift on a 0–100 scale — safety comes from the anchor pins and the batch structure, not from catching every adversarial contribution (Table 1, REPORT.md).

Figure 2. Median drift on the released anchors, round 3, EN, $p_{\text{band}} = 0.7$, $f_{\text{exp}} = 0.2$, all three contribution volumes. Every grid point sits under 0.04 FRS points, far below the 0.5-point line at which drift would become measurement-relevant on the 0–100 scale.

Detection. The quarantine gate intercepts roughly half of adversarial contributions (English: precision .53, recall .55; Dutch: precision .41, recall .58), at a comparable false-positive cost on honest-but-noisy input. The gate cannot distinguish wrong from malicious, by construction; the safety result shows it does not need to.

Benefit. Improvement in held-out placement accuracy materializes only when label quality is decent: with correct-band probability at 0.7 or better and a meaningful expert fraction, rounds improve the scale; at 1,000 low-quality contributions per round, three rounds fail to clear a one-FRS-point benefit threshold. Volume does not substitute for quality. This is the empirical vindication of the intake design of §3: the anchored question, the typed labels, and the vote dosage rule exist precisely to raise the quality term that the simulation shows is the binding constraint.

Improvement tracks label quality, not contribution volumeround-3 held-out MAE improvement, $EN, f_{exp} = 0.2$, adversarial = 0

At 50 contributions/round every quality level is noise-dominated and negative. At 1000, the lines separate cleanly by label quality: the 0.9-accuracy line reaches +0.10 FRS while 0.5-accuracy stalls at +0.01 — volume alone does not help.

Figure 3. Round-3 held-out MAE improvement against contribution volume, $EN, f_{exp} = 0.2$, adversarial = 0, one line per band-label accuracy. At 50 contributions per round every quality level is noise-dominated and negative; at 1,000, the lines separate by label quality rather than by volume, and only the higher-quality conditions cross into positive territory.

A methods by-product. Joint Bradley-Terry refits that treat single-vote contributions as free parameters are numerically unstable (pendant nodes in the comparison graph), an instability confirmed to exist in the production fitter beyond its shipped iteration count. The round algorithm's fixed-position entry for low-vote contributions, and the ≥ 10 -vote dosage rule, are the mitigations.

5.3 What the simulation does not establish

Synthetic contributors follow parameterized noise models; real contributors may be stranger. The adversarial model is generic (inversion and randomness), not targeted. The simulation does not model contributor volume dynamics, seasonality, or the correlation structure of real texts people choose to submit. These are precisely the things the first operational rounds will reveal, and the pre-registered comparison of round-one reality against this study is reserved for the follow-up paper.

6. Discussion

Why batched, one more time. Every result above depends on the batch structure. A continuously updating scale would expose users to drift as it happens, make adversarial probing interactive, and reduce every reliability claim to a moving target. Batching converts mainten-

ance into discrete, gate-checked, publicly changelogs events, which is what allows a free public intake to coexist with a stable standard.

Governance is part of measurement. The round algorithm ends in a human decision: gates green, change notes written, version shipped. The accountability mechanism is publication itself; a round's full inputs and decisions are inspectable, which no proprietary scale offers.

Two data streams, one architecture. The site intake is the first stream. The second, described in the companion education paper (Vande Moortel, 2026c), is classroom telemetry: consented oral-reading outcomes from schools using the FRS Teams application, which would ground the reader side of the scale. Both streams terminate in the same round algorithm; the intake design differs, the maintenance discipline does not.

7. Limitations

Simulation-only evidence for the round algorithm; the first real round is the actual test and is pre-committed to publication either way. The quarantine gate's detection ceiling is modest and inherently confounds error with malice. The Dutch mapping's control-point non-monotonicity, worked around during the simulation, needs a structural fix in the calibration scripts. The expert-tolerance claim from the comparative-judgment literature rests partly on sources that resisted primary retrieval and is used only in weakened form. The novelty claim for pooling a language-model judge with human raters in one rater-quality model awaits one final dedicated literature pass.

8. Conclusion

An open standard does not have to choose between freezing and drifting. With an intake that asks people questions they can actually answer, a rater-quality model that earns trust instead of assuming it, anchors that are pinned and monitored rather than trusted blindly, and rounds that ship like software releases, a reading-difficulty scale can grow from public contribution while keeping the one promise a standard must keep: the same text, the same version, the same score. The intake is live at the standard's public site. Round one will say whether the simulation was a good prophet; the design commits us to publishing the answer.

References

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment* (Cambridge Assessment Research Report). Cambridge Assessment. [Figures cited via secondary reports; primary PDF not machine-readable at time of writing.]

- Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(1), 43–58.
- Chen, X., Bennett, P. N., Collins-Thompson, K., & Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*, 193–202.
- Crossley, S. A., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55, 491–507.
- Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1), 20–28.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
- Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281–300.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286.
- Vande Moortel, J. (2026a). *The Free Reading Standard: Definition and construction of an open, continuous scale for technical reading difficulty in Dutch and English* (vo.2). Orion-Solution. <https://frs.orion-solution.net/research>
- Vande Moortel, J. (2026b). *Calibrating an open reading-difficulty scale without human-labeled corpora* (vo.1). Orion-Solution. <https://frs.orion-solution.net/research>
- Vande Moortel, J. (2026c). *Measuring reading growth: FRS and Microsoft Reading Progress in the classroom* (in preparation). Orion-Solution.
- Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5), 541–562.
- Bracketed notes mark details resting on secondary sources; they are resolved before any published version.*

Appendix A. Intake form copy (verbatim, NL + EN)

Quoted verbatim from the deployed site's locale files (`site/src/i18n/locales/{nl,en}.json`, keys under `analyzer.contribute`), the source of truth for the shipped copy. The structural specification is §3; the visual flow is Figure 4.

Intro. NL: "FRS verbetert via gebundelde herkalibratierondes op basis van echte oordelen, niet alleen ankerteksten. Drie korte, optionele stappen — geen enkele vraagt je om zelf een cijfer op de schaal te plakken." EN: "FRS improves through batched recalibration rounds fed by real judgments, not just anchor texts. Three short, optional steps — none of them ask you to paste a number onto the scale yourself."

Step 1 — audience band (everyone). NL question: "Wie kan deze tekst vlot lezen, denk je?" EN: "Who could read this text comfortably?" Placeholder NL: "Kies een doelgroep..." / EN: "Pick an audience..."

Options (NL / EN):

- eerste leerjaren ($\pm 6-8$ j) / early primary (age $\sim 6-8$)
- derde–vierde leerjaar ($\pm 8-10$ j) / middle primary (age $\sim 8-10$)
- vijfde–zesde leerjaar ($\pm 10-12$ j) / upper primary (age $\sim 10-12$)
- secundair / secondary school
- volwassenen, algemeen / adults, general
- gespecialiseerd/academisch / specialist/academic
- weet ik niet / not sure

Step 2 — typed expert label (optional, disclosed). Toggle NL: "Ik ken het niveau van deze tekst" / EN: "I know this text's level." Fields, NL / EN:

- "AVI-niveau (voor leerkrachten)" / "AVI level (for Dutch-language teachers)", placeholder "Kies AVI-niveau..." / "Pick an AVI level..."
- "CEFR-niveau (voor taaldocenten/NT2)" / "CEFR level (for language teachers/NT2)", placeholder "Kies CEFR-niveau..." / "Pick a CEFR level..."
- "Waar komt dit niveau vandaan?" / "Where does this level come from?", placeholder "bv. boekomslag, leesmethode, eigen inschatting" / "e.g. book cover, reading method, own assessment"

Role and source (recorded with every contribution). Role label NL: "Wie geeft dit oordeel?" / EN: "Who is giving this judgment?" Options: Leerkracht/Teacher, Ouder/Parent, Auteur/Author, Andere/Other. Source label NL: "Waar komt deze tekst vandaan?" / EN: "Where does this text come from?" Options: Eigen tekst/My own text, Schoolboek/Textbook, Website/Website, Andere/Other.

Consent and contact. Email label NL: "E-mail (optioneel, enkel voor een antwoord)" / EN: "Email (optional, only if you want a reply)." Consent label NL: "Ik ga ermee akkoord dat dit oordeel en de tekst hierboven bewaard en gebruikt mogen worden voor FRS-kalibratieonderzoek." / EN: "I agree that this judgment and the text above may be stored and used for FRS calibration research." Submit button NL: "Bijdrage versturen" / EN: "Send contribution." Success message NL: "Bedankt — je bijdrage is geregistreerd." / EN: "Thank you — your contribution was recorded."

Step 3 — comparison vote (everyone, the gold signal). Title NL: "Nog één ding: welke tekst is moeilijker?" / EN: "One more thing: which text is harder?" Intro NL: "Een korte vergelijking is het meest waardevolle oordeel dat je kan geven — dit voedt rechtstreeks de volgende kalibratieronde." / EN: "A short comparison is the most valuable judgment you can

give — it feeds directly into the next calibration round." Loading state NL: "Vergelijkingsfragment laden..." / EN: "Loading a comparison excerpt..."

Panel labels NL / EN: "Jouw tekst" / "Your text"; "Dit fragment" / "This excerpt." Question NL: "Welke tekst is moeilijker om te lezen?" / EN: "Which text is harder to read?" Choice buttons NL: "Jouw tekst is moeilijker" / "Dit fragment is moeilijker" — EN: "Your text is harder" / "This excerpt is harder." Thanks message NL: "Bedankt voor je vergelijking!" / EN: "Thanks for the comparison!"

Appendix B. Round algorithm pseudocode

Generated from the reference implementation's simulation library (`calibration/experiments_flywheel/common.py`), which reuses the production Bradley-Terry fitter (`03_fit_bt.py`) and ridge pipeline factory (`05_fit_ridge.py`) rather than re-deriving them. Numbered to match §4's six steps; constants are the module's own, not invented for this appendix.

```

INPUT:  current scale version S, contribution pool P, anchor set A (pinned)
OUTPUT: new scale version S', change notes

# 1. ELIGIBILITY
for each contribution c in P:
    # dosage rule (>= 10 votes, §2.1); provenance required for labels
    c.vote_eligible = (c.accumulated_votes >= 10)
    c.label_eligible = (c.role is set AND c.source is set)
    if not (c.consented and not c.quarantined_at_intake):
        drop c from this round

# 2. QUARANTINE GATE (quarantine_flag)
for each remaining contribution c:
    # expert label if given, else band midpoint, nudged by its own vote
    label_est = label_estimate_frs(c)          # +/- VOTE_NUDGE_FRS (2.5)
    model_pred = frozen_predict(S, c.features) # current model's estimate
    # QUARANTINE_MODEL_DISAGREE_FRS = 25.0
    disagree_model = |model_pred - label_est| > 25.0
    gap = label_est - mapped_frs[c.anchor_id]
    # QUARANTINE_VOTE_MISMATCH_FRS = 10.0
    dir_mismatch = (|gap| > 10.0) AND (sign(gap) != c.vote_says_harder)
    c.quarantined = disagree_model OR dir_mismatch
    # quarantine flags for review; it does not delete c

passed = [c for c in remaining if not c.quarantined]

# 3. RATER-QUALITY WEIGHTING
for each c in passed:
    # W_EXPERT_LABEL = 3.0, W_BAND_LABEL = 1.0
    c.role_weight = 3.0 if c.is_expert_label else 1.0
    # the LM judge that produced S's own anchor fit is itself one more
    # rater in the same pool, weighted on the same footing (§2.3)

# 4. BLENDED REFIT WITH PINNED ANCHORS
vote_sample = subsample(passed, cap=300, seed=fixed) # BT_NODE_CAP
# anchor bt_score[A] is NOT recomputed here: A stays exactly as
# published in S. only the feature-to-scale regression sees the
# round's new evidence:
target[a] = mapped_frs_S[a]          for each anchor a in A # pinned
target[c] = label_estimate_frs(c)    for each surviving contribution c
ridge_S' = fit(
    StandardScaler + Ridge(alpha=S.alpha),
    features[A + surviving_contributions],
    target[A + surviving_contributions]) # per language

# 5. DRIFT MONITOR
# (blend_anchor_bt, pins released -- shadow fit only, never shipped)
s = anchor_strength0[A] # S's own fit, already at its stopping point
s = s / mean(s)        # (not re-derived)
# max_iter=3, NOT run to convergence: production's own stopping
# convention; more iterations measurably destabilize this
# ill-conditioned graph (Appendix C)
repeat up to 3 times:

```

```

for each anchor a in A:
  N[a,*] = base_games[a,*]          # real judgment graph
  extra_wins[a] += sum(weight
    for (anchor_id==a, contribution_harder==False)
    in vote_sample)
  contrib_str[a] += sum(
    weight / (s[a] + exp(clip(inverse_frs_map(label_est),
                              -100, 100))))
    for each vote touching a)
  s = MM_update(s, N, extra_wins, contrib_str) # BT MM step
  s = s / mean(s)                             # renormalize, as S was
shadow_bt[A] = log(s)
drift[a] = |shadow_bt[a] - bt_score_S[a]| for each anchor a
# median is the trusted statistic (§5.2); max is dominated by
# near-zero-strength anchors unstable under ANY finite fit
report median(drift), mean(drift), p90(drift), max(drift)
flag anchors where drift[a] exceeds the published tolerance
  # flagged anchors are retired or re-labeled in a LATER round,
  # never moved silently here

# 6. GATES AND RELEASE
run full calibration validation-gate suite against S'
if any gate fails:
  ABORT -- S ships unchanged, failure reasons logged
else:
  S' <- { ridge_S',
    bt_score[A] = bt_score_S[A], # still pinned
    version = S.version + 1 }
  publish change notes:
    { n_entered, n_quarantined, anchors_flagged, drift_summary }
  FREEZE S' until the next round

```

Appendix C. Simulation grid and artifacts

Full condition grid, per-condition tables, and reproducibility statement: `calibration/experiments_flywheel/` in the reference implementation (`REPORT.md`, `sim_results.json`, `sim_raw.json`; two full-grid executions byte-identical).

Change notes v0.1: first draft, written before the mechanism's first operational round by design. Figures planned for v0.2: the flywheel loop, the safety plot (drift vs adversarial fraction), the quality-not-volume plot, the intake flow.

Change notes v0.2: figures added; Appendix B pseudocode generated from implementation.