

# The Free Reading Standard: Definition and Construction of an Open, Continuous Scale for Technical Reading Difficulty in Dutch and English

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

---

## Abstract

Placing a text on a reading difficulty scale is a solved problem only for those who can pay for it. The dominant instruments are either discrete and closed (the Dutch AVI system assigns one of twelve levels through a paid, publisher-facing service) or continuous but proprietary (the Lexile Framework and ATOS publish their scale semantics while withholding their current production formulas). For Dutch, no open continuous scale exists at all; no open instrument covers Dutch and English on comparable footing. This paper defines the Free Reading Standard (FRS): an open specification for measuring the technical reading difficulty of a text on a continuous 0–100 scale spanning the full spectrum of written material, from a child's first decodable words to dense academic prose, reported as a four-dimension profile (decodability, lexical difficulty, syntactic complexity, referential cohesion) plus a composite score with an explicit uncertainty band. Calibration combines anchor texts carrying trusted difficulty labels with pairwise difficulty judgments produced by a large language model and aggregated with a Bradley-Terry model, a paradigm validated with human judges by the CLEAR corpus project. Two pilot experiments (20 Dutch texts, 780 model calls) motivate the design: absolute rubric scoring by an LLM was reliable ( $ICC = .97$ ) but compressed texts into a quarter of the scale, while pairwise judgment nearly halved ordering errors against reference labels (13 versus 24 violations of 190 pairs). All parameters, prompts, word lists, and anchor procedures are published. The standard is a specification and a research program: reader-side validation is future work, and is stated as such.

*Keywords:* readability, text difficulty, technical reading, Bradley-Terry, large language models, Dutch, open standards

---

## 1. Introduction

A teacher who wants to know whether a text suits a reader has, in the Dutch-speaking world, essentially one instrument: the AVI level printed on the back of a children's book. That level is one of twelve discrete steps, it stops discriminating entirely somewhere around age eleven, and the procedure that assigns it is a commercial service: publishers submit books to Cito and pay for the assessment and the right to print the logo (see §2.1). In the English-speaking world the

situation is better funded but structurally similar. The Lexile Framework offers a genuinely continuous, psychometrically constructed scale (Stenner, 1996; Stenner et al., 2007), but its current production formula is proprietary, and the same holds for ATOS (School Renaissance Institute, 2000). A school, a tool builder, or a researcher who wants to place an arbitrary text on a fine-grained difficulty scale, reproducibly and at no cost, currently cannot do so in Dutch and can do so in English only with crude surface formulas from the 1940s to 1970s whose limitations are well documented (see §2.2).

This paper defines the Free Reading Standard (FRS), an open specification for the technical reading difficulty of texts. Technical reading difficulty means the demand a text places on decoding and fluent oral reading: how hard the words and sentences are to read, not how hard the ideas are to understand. This is the construct the AVI tradition measures on the student side through timed oral reading (accuracy and speed), and deliberately not reading comprehension, which involves distinct reader and text factors and is planned as a separate future extension of the standard.

The choice of construct has a practical origin. Microsoft's Reading Progress and Reading Coach tools, deployed in schools through Teams, produce exactly these measurements (accuracy, words per minute, error categories) at scale, but they lack any principled way to characterize the difficulty of the text that was read. Without a text-side measure, a reader's 97% accuracy at 80 words per minute is uninterpretable: excellent on a hard text, worrying on an easy one. FRS exists to supply the text side of that equation, openly.

The contributions of this paper are:

1. **An open specification** for a continuous technical reading difficulty scale covering Dutch and English: construct definition, a four-dimension difficulty profile with per-language feature sets, a composite score on a 0–100 scale spanning the full spectrum of written text with mandatory uncertainty reporting, and anchor semantics tied to existing practice (a published AVI mapping for Dutch; graded corpora for English).
2. **A calibration method** that scales beyond scarce labeled data: pairwise difficulty judgments produced by a large language model under a fixed protocol, aggregated with a Bradley-Terry model, and mapped onto the anchored scale by monotone calibration. The pairwise-comparison paradigm for readability was established with human judges by the CLEAR corpus project (Crossley et al., 2023); to our knowledge its combination with an LLM judge, and its application to Dutch, are both new.
3. **Radical openness as a design requirement.** Every coefficient, prompt, word list, anchor list, and versioning rule is published (Appendices A–D). Where comparable systems publish their scale but not their formula, FRS treats reproducibility as part of the construct: a standard nobody can re-implement is a product, not a standard.
4. **Preliminary empirical motivation.** Two pilot experiments on 20 Dutch texts with reference labels show that the failure mode of direct LLM scoring is scale compression

rather than unreliability, and that pairwise judgment materially improves ordering quality (§4). These pilots motivate the design; they do not validate the standard. Validation at scale is the subject of a companion paper in preparation.

The remainder of the paper reviews what exists and why it is insufficient (§2), specifies the standard (§3), reports the pilots (§4), and discusses positioning, risks, and limitations (§5–6).

---

## 2. Background and related work

### 2.1 Discrete leveling: the AVI tradition

The AVI system (originally *Analyse van Individualiseringsvormen*), introduced in 1972 and overhauled in 2008, is the de facto standard for technical reading levels in the Netherlands and Flanders. It defines twelve levels (Start, M3, E3, M4 ... E7, Plus), indexed to the middle (M) and end (E) of primary school years. On the text side, a book's level is determined by the CILT index (*Cito Index voor de LeesTechniek*), which uses two variables: the percentage of high-frequency words, per the frequency list of Staphorsius, Krom and De Geus (1988), and mean word length in letters. We verified the published coefficients against Cito's own technical documentation and an independent refit on the thirty published AVI card texts ( $U = 114.49 + 0.28 \cdot \% \text{high-frequency} - 12.33 \cdot \text{letters per word}$ ;  $\text{CILT} = 150 - U$ ; lower CILT values correspond to easier texts, with level boundaries running from  $\leq 56.9$  for AVI-M3 to  $> 74.9$  for AVI-Plus). Notably, the 2008 revision dropped sentence length as a predictor for the technical reading index. On the student side, level placement uses timed oral reading of standard cards (accuracy and speed thresholds per level) alongside the DMT word-reading test.

Three properties of this system motivate FRS. First, granularity: twelve bins, with nothing above AVI-Plus, so the system cannot express within-level growth or any difficulty differences relevant beyond age eleven. Second, closedness: text leveling is a paid, publisher-facing service, and the operational details (such as the exact high-frequency cutoff applied to the 1988 frequency list) are not published in full. Third, norm fragility: the levels travel worse than practitioners assume. When Flemish researchers re-normed the DMT and AVI cards on 1,461 Flemish pupils, Flemish children scored significantly lower than the Dutch norm group on the DMT in all six primary grades, and read the AVI cards significantly more slowly in every grade but one (Leysen et al., 2018). The same instrument, the same language, one border, different norms. A difficulty standard should therefore separate the text-side scale (which can be universal) from reader-side norms (which are demonstrably local).

### 2.2 Readability formulas, from Flesch to CLIB

The classic formula tradition needs only brief treatment. Flesch (1948) and the Flesch-Kincaid grade-level variant (Kincaid et al., 1975) regress comprehension criteria on two surface

variables (sentence length and syllables per word); they were adopted far beyond their validation scope and carry no frequency or semantic information. Dutch adaptations followed the same pattern: Douma's adjustment of Flesch's coefficients (published 1960, per the archived documentation of the Kennisbank Begrijpelijke Taal) and Brouwer's Leesindex A, published in 1963 with per-word coefficients and rescaled to per-100-words units in Brouwer's 1974 extension to the B and AB variants; the two statements are algebraically the same formula (Brouwer, 1963).

The serious Dutch contribution is CLIB (Staphorsius, 1994), the *Cito Leesbaarheidsindex voor het Basisonderwijs*, a comprehension-oriented readability index built on four variables (mean word length, percentage of high-frequency words, type-token ratio, and sentences per word). Its formula-side validation was substantial for its era: 240 texts, 4,320 pupils, and 8,640 cloze tests, with reported multiple correlations around .85 at the text level. More important for FRS is what CLIB became institutionally. Cito's scientific documentation for its pupil-monitoring comprehension tests (Feenstra et al., 2010) shows that CLIB functions as a conjoint scale in the psychometric sense: item banks are calibrated with the One-Parameter Logistic Model, and both pupil ability and text difficulty are expressed as values on one common scale. The Netherlands, in other words, already possesses a structural analogue of the Lexile Framework for Dutch. It is closed, embedded in a commercial test suite, and unavailable as a general-purpose text measurement service. FRS's ambition is precisely this architecture, in the open.

## 2.3 Continuous calibrated scales: the Lexile blueprint

The Lexile Framework remains the reference construction for a continuous reading scale (Stenner, 1996; Stenner et al., 2007). Its essential move is conjoint measurement: reader ability and text difficulty are placed on the same interval scale, in logits, through Rasch modeling of real students answering items tied to real passages. The text side is predicted from two variables, log mean sentence length and mean log word frequency over a corpus of roughly 600 million words. The original specification equation is public (Stenner, 1996):  $\text{Logit} = 9.82247 \cdot \log \text{mean sentence length} - 2.14634 \cdot \text{mean log word frequency} - \text{constant}$ , rescaled as  $\text{Lexile} = (\text{Logit} + 3.3) \times 180 + 200$ , anchored so that basal primers sit near 200L and grade-12 encyclopedic text near 1200L, with the scale's semantics fixed by a 75% comprehension criterion. Reported correlations between formula-predicted and empirically calibrated text difficulties reach .92 uncorrected (.97 corrected) on a single test battery and .93 corrected across nine batteries (Stenner et al., 2007). Two lessons carry over to FRS. First, a formula becomes a scale only through anchoring to something with fixed meaning. Second, openness is a choice: the original Lexile equation is citable and replicable, yet the production formula in commercial use today is not, and ATOS's coefficients never were (School Renaissance Institute, 2000).

## 2.4 One number is not enough

A parallel research tradition holds that a single difficulty number, however well calibrated, hides the information users actually need. ETS's TextEvaluator reports eight separate dimensions of text complexity alongside its overall placement (Sheehan et al., 2015); the Coh-Metrix Text Easability Assessor reports five (narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion). The Common Core State Standards' guidance institutionalized the same caution, requiring quantitative measures to be combined with qualitative judgment and reader-task considerations (National Governors Association & CCSSO, 2010). The practical failure case is well documented: two texts with identical scores and very different actual difficulty profiles.

The Dutch feature-selection literature sharpens this into a concrete short list. Kleijn (2018), in the dissertation underlying the LiNT instrument, screened more than 400 automatically extracted text features against cloze-measured comprehension for 2,926 secondary school pupils, using a purpose-built rational-deletion cloze methodology with demonstrated reliability (mean Cronbach's  $\alpha = .83$  for semantic scoring). Five features survived stepwise selection: content-word frequency (by far the strongest,  $\Delta R^2 = .173$ ), content words per clause, concrete nouns, maximal syntactic dependency length, and adjectival past participles; the deployed LiNT instrument uses the top four. Text features jointly explained about 23% of comprehension variance, against roughly 34% for reader characteristics, a proportion worth keeping in mind whenever any text-side instrument, including FRS, is tempted to overclaim. The same dissertation found that this feature set outperformed both the Flesch-Douma and CLIB predictor sets re-estimated on the same data.

## 2.5 Decodability: why Dutch and English need different features

Because FRS targets technical reading, its lowest-level dimension concerns decodability: the difficulty of converting print to sound. The psycholinguistic evidence is firm that this difficulty is orthography-dependent. In the landmark cross-language comparison of foundation literacy across thirteen European orthographies, most children reached accurate and fluent foundation reading within the first school year; English was the outlier, attributable to two separable factors, orthographic depth and syllabic complexity (Seymour, Aro, & Erskine, 2003). Grain-size theory explains the mechanism: readers of consistent orthographies like Dutch can rely on small-grain letter-sound mappings, while English forces mixed strategies over larger units (Ziegler & Goswami, 2005). Word-length and cluster effects compound this at the item level, and consonant clusters in particular amplify length effects for beginning readers.

For the specification this has one blunt consequence: a shared cross-language decodability formula is not defensible. The Dutch feature set can lean on the structural progression institutionalized in Dutch initial reading instruction (the *klankzuiver* criterion and the consonant-vowel structure taxonomy of the Veilig Leren Lezen tradition), while the English set must

additionally weight grapheme-phoneme irregularity, for which Hiebert's TExT model offers a documented precedent, separating linguistic content (vowel-pattern complexity of monosyllabic words) from cognitive load (unique-word counts) (Hiebert, 2024). We also note, as a scope discipline, the caution running through the decodable-text research, synthesized in the practitioner literature most prominently by Shanahan, against treating decodability as a pedagogical prescription rather than a measurement axis: FRS measures the demand; it does not prescribe instruction.

## 2.6 Machine learning, LLMs, and the consistency problem

Automatic readability assessment has moved through three architectures: handcrafted features with classical models, fine-tuned transformers, and prompted large language models. Two findings from this literature shaped FRS. First, hybrid systems that combine handcrafted linguistic features with neural representations outperform either alone, with the advantage largest on small datasets (Lee, Jang, & Lee, 2021; see Vajjala, 2021 for a survey of the field's persistent evaluation problems). This matters because the Dutch labeled-data situation is small-data by construction (§2.7). Second, prompted LLMs are usable but imperfect absolute judges of readability: zero-shot GPT-4-class estimates correlate with human judgments at  $r \approx .74-.76$ , beating classic formulas but leaving substantial variance unexplained, with the authors cautioning that readability is context- and audience-dependent (Trott & Rivière, 2024).

The alternative to absolute judgment is relative judgment. The CLEAR corpus (Crossley et al., 2023) demonstrated the paradigm at scale for readability: 1,116 quality-filtered raters produced 111,347 pairwise "which passage is easier" judgments over roughly 5,000 excerpts, aggregated with a Bradley-Terry model into continuous difficulty scores (final corpus 4,724 excerpts after quality exclusions; CC BY-NC-SA 4.0). The method traces to earlier crowd-sourced readability work (De Clercq et al., 2014, as cited in Crossley et al., 2023). Independent support for the relative framing comes from fine-tuned pairwise rankers, which achieve near-ceiling pairwise accuracy on graded corpora where absolute classification remains hard (Lee & Lee, 2023). What the literature does not yet contain, to our knowledge, is this paradigm executed with an LLM as the judge, nor any psychometrically anchored continuous readability scale covering Dutch: the only multilingual system including Dutch is a zero-shot ranking model without calibrated scale semantics (Trokhymovych, Sen, & Gerlach, 2024).

## 2.7 The data landscape and the gap

The asymmetry between the two target languages is stark. English offers CLEAR (continuous labels, open), OneStopEnglish (three aligned levels, open; Vajjala & Lučić, 2018), and several request-gated corpora. Dutch offers no open graded corpus at all: AVI labels live in proprietary publisher and library metadata; the graded news products are subscription services; and the open resources are either frequency norms (SUBTLEX-NL, Keuleers, Brysbaert, & New, 2010;

word prevalence norms, Keuleers et al., 2015; the child-directed BasiLex lexicon) or simplification pairs without difficulty labels. Perhaps one to fifteen thousand genuinely difficulty-labeled Dutch text units exist across scattered, mostly closed sources. This scarcity is not an inconvenience to be regretted; it is the design constraint the calibration method of §3.4 is built around, and the reason the method is validated on English (where gold standards exist) before being trusted for Dutch.

The gap FRS fills can now be stated precisely (Figure 1). There exists no open, reproducible, continuous scale for technical reading difficulty; none for Dutch of any kind (the closed CLIB conjoint scale being the nearest analogue); no instrument covering Dutch and English with common semantics; and no published application of the pairwise-aggregation paradigm with an LLM judge. FRS claims exactly these four novelties, and no others.

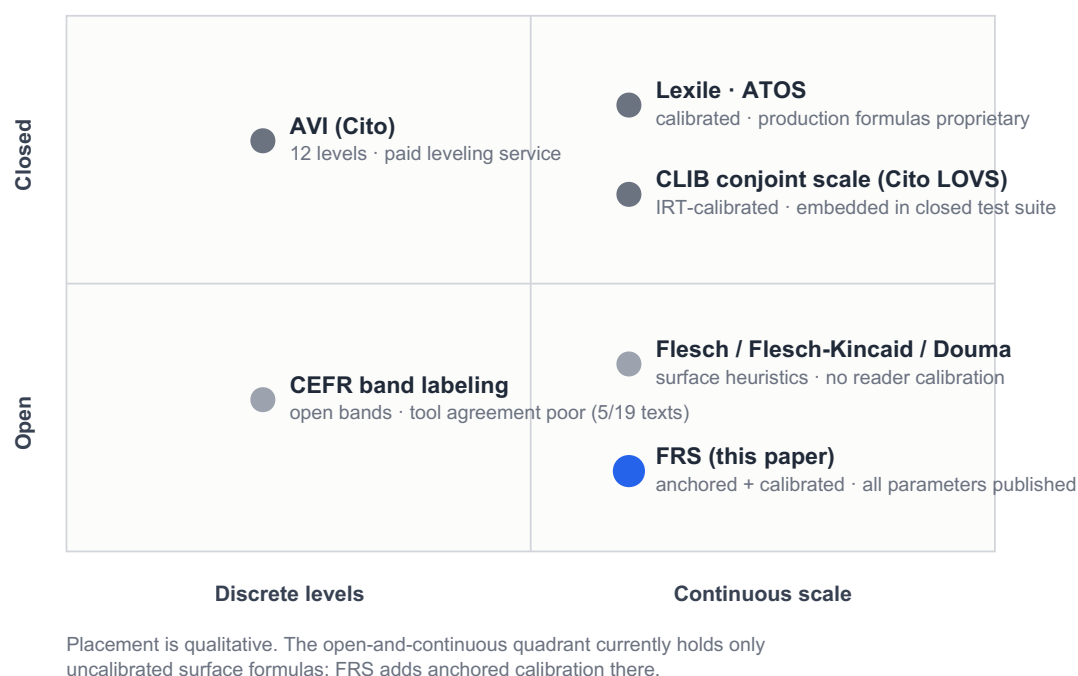


Figure 1. The instrument landscape by openness and granularity. The open-and-continuous quadrant currently holds only uncalibrated surface formulas; FRS adds anchored calibration there.

### 3. The Free Reading Standard: specification

#### 3.1 Construct

FRS measures the technical reading difficulty of a text: the demand it places on accurate and fluent oral reading. Operationally, a text is technically harder than another if a population of readers, matched on ability, reads it with lower accuracy and lower speed. The construct explicitly excludes comprehension difficulty, background-knowledge demands, and interest or

motivation, which belong to a planned comprehension extension. The scale covers Dutch and English. The two languages share the architecture, the output format, and the scale semantics; they do not share feature formulas (per §2.5) or calibration data. Whether the two calibrations can be linked into one cross-language scale, such that equal FRS scores imply comparable difficulty across languages, is an open research question and is not claimed.

### 3.2 The difficulty profile

Every analyzed text receives four sub-scores, each on a published 0–100 scale, each computable deterministically from the text alone, each explainable to a teacher in one sentence:

| Dimension            | One-sentence meaning                             | Core features (Dutch)                                                                                                       | Core features (English)                                                                                                                              |
|----------------------|--------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Decodability         | How hard the words are to sound out              | word length in letters and phonemes; consonant-cluster complexity; conformity to the klankzuiver criterion; digraph density | as Dutch, plus grapheme-phoneme irregularity weighting (vowel-pattern complexity of monosyllabic words)                                              |
| Lexical difficulty   | How rare or unfamiliar the words are             | SUBTLEX-NL content-word frequency (compound-corrected, per Kleijn 2018); word prevalence; age-group frequency (BasiLex)     | SUBTLEX-US/UK frequency; word prevalence (Brysbaert et al., 2019); age-of-acquisition (Kuperman et al., 2012); concreteness (Brysbaert et al., 2014) |
| Syntactic complexity | How much sentence structure the reader must hold | mean sentence length; maximal dependency length; content words per clause                                                   | same feature definitions, English parser                                                                                                             |
| Referential cohesion | How much the text helps the reader connect ideas | argument overlap; connective density; concrete-noun ratio                                                                   | same feature definitions, English resources                                                                                                          |

The feature sets deliberately start from the empirically survived Dutch short list (Kleijn, 2018) extended with the decodability features that comprehension-oriented instruments omit but technical reading requires (Figure 4). Full operational definitions, resources, and versioned parameter values are given in Appendix A. Feature extraction is deterministic: the same text and the same FRS version always produce identical feature values.

### The FRS difficulty profile: four dimensions, per-language features

| Dimension                                                              | Dutch features                                                                                              | English features                                                                 |
|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| <b>Decodability</b><br>how hard the words are to sound out             | word length (letters, phonemes)<br>consonant-cluster complexity<br>klankzuiver conformity · digraph density | as Dutch, plus:<br>grapheme-phoneme irregularity<br>vowel-pattern complexity     |
| <b>Lexical difficulty</b><br>how rare or unfamiliar the words are      | SUBTLEX-NL frequency (content words)<br>word prevalence (Ghent norms)<br>per-grade frequency (BasiLex)      | SUBTLEX-US/UK frequency<br>prevalence · age of acquisition<br>concreteness norms |
| <b>Syntactic complexity</b><br>how much structure the reader must hold | mean sentence length<br>maximal dependency length<br>content words per clause                               | same feature definitions,<br>English parser and resources                        |
| <b>Referential cohesion</b><br>how much the text helps connect ideas   | argument overlap<br>connective density<br>concrete-noun ratio                                               | same feature definitions,<br>English parser and resources                        |

Feature short list grounded in the Dutch feature-survival evidence (Kleijn, 2018) plus decodability features required by the technical-reading construct; decodability differs per orthography (Seymour et al., 2003).

Figure 4. The four FRS dimensions with their per-language feature sets. Decodability features differ per orthography; the higher dimensions share definitions across languages.

### 3.3 The composite score

The headline output is one continuous score on the FRS 0–100 scale, always reported with an uncertainty band (for example, FRS 34 ± 3), never as a bare point value. The scale is designed for the full spectrum of written material: a child's first decodable word lists and rhymes near the bottom, everyday news prose in the middle regions, and dense specialist or academic prose near the top, with headroom above the hardest current anchors so the scale never saturates the way the discrete AVI range does. Its semantics are fixed entirely by published anchors, not by the number 100. (An earlier FRS prototype used a 0–12 rubric; that scale is retired, and the pilots of §4 that used it are reported in its own terms.)

For Dutch, the standard publishes a mapping table between the AVI region of the spectrum and AVI levels, so that practice anchored in AVI transfers. Indicatively, the discrete AVI range (Start through Plus) occupies roughly the bottom third of the FRS scale, with the E7/Plus boundary as a named anchor point; the exact mapping values are fixed during calibration and published with each FRS version, and the table below illustrates the intended form rather than final values.

| AVI level | FRS region (illustrative, fixed at calibration) |
|-----------|-------------------------------------------------|
| Start     | ≈ 2                                             |
| M3–E3     | ≈ 5–9                                           |

| AVI level       | FRS region (illustrative, fixed at calibration) |
|-----------------|-------------------------------------------------|
| M4–E5           | ≈ 10–18                                         |
| M6–E7           | ≈ 19–28                                         |
| Plus (boundary) | ≈ 30                                            |

Anchor semantics for English are set by graded-corpus reference points at equivalent stages (early decodable readers through academic text). The composite is not a hand-weighted sum of the profile: it is fit during calibration (§3.4), and the profile exists so that identical composites with different difficulty sources remain distinguishable, answering the "same score, different text" failure of single-number instruments (§2.4).

### 3.4 Calibration: anchors, pairwise LLM judgments, Bradley-Terry

The calibration problem is to place arbitrary texts on the anchored scale when labeled data is scarce (Dutch) or license-restricted (English). FRS solves it in three steps.

**Anchor sets.** A published list of anchor texts with trusted difficulty positions: for Dutch, texts with verified AVI card levels at the low end and staatsexamen-style CEFR-graded material above; for English, a stratified sample spanning the CLEAR difficulty range. Anchor lists, provenance, and label verification status are published (Appendix D). Anchor quality is a first-class concern: our pilot surfaced one probable label error among twenty seed texts (§4.3), and every anchor therefore carries a provenance record.

**Pairwise judgment protocol.** An instructed large language model judges, for pairs of texts, which is technically harder to read. The protocol is fixed and versioned: model identifier pinned, temperature 0, the exact system prompt of Appendix B, texts truncated to a fixed window, and, critically, every pair judged in both presentation orders. The both-orders rule is not a nicety: in our pilot, 16% of pairs flipped their verdict when presentation order was reversed (§4.2). Order effects are a documented failure mode of LLM judgment, and single-presentation designs silently absorb them into the scale.

**Aggregation and mapping.** Judgments are aggregated with a Bradley-Terry model, yielding a latent difficulty value per text, exactly as in the human-judge construction of the CLEAR corpus (Crossley et al., 2023). The latent scale is then mapped onto the anchored 0–100 scale by monotone calibration through the anchor texts. The uncertainty band of §3.3 derives from the Bradley-Terry standard errors propagated through this mapping. Because the deterministic feature core (§3.2) and the fitted mapping are frozen at release, scoring a new text does not require fresh LLM judgments in the common case: the released scale model predicts placement from features, with pairwise judgment reserved for calibration rounds, disputed texts, and scale maintenance.

**Versioning.** Any change to the judge model, the prompt, the anchor list, or the feature definitions produces a new FRS version with published change notes and re-validation. Version identifiers accompany every score. An FRS score without a version is meaningless, and the specification says so. Figure 2 summarizes the full architecture, including the contribution loop of §7 that feeds batched recalibration rounds.

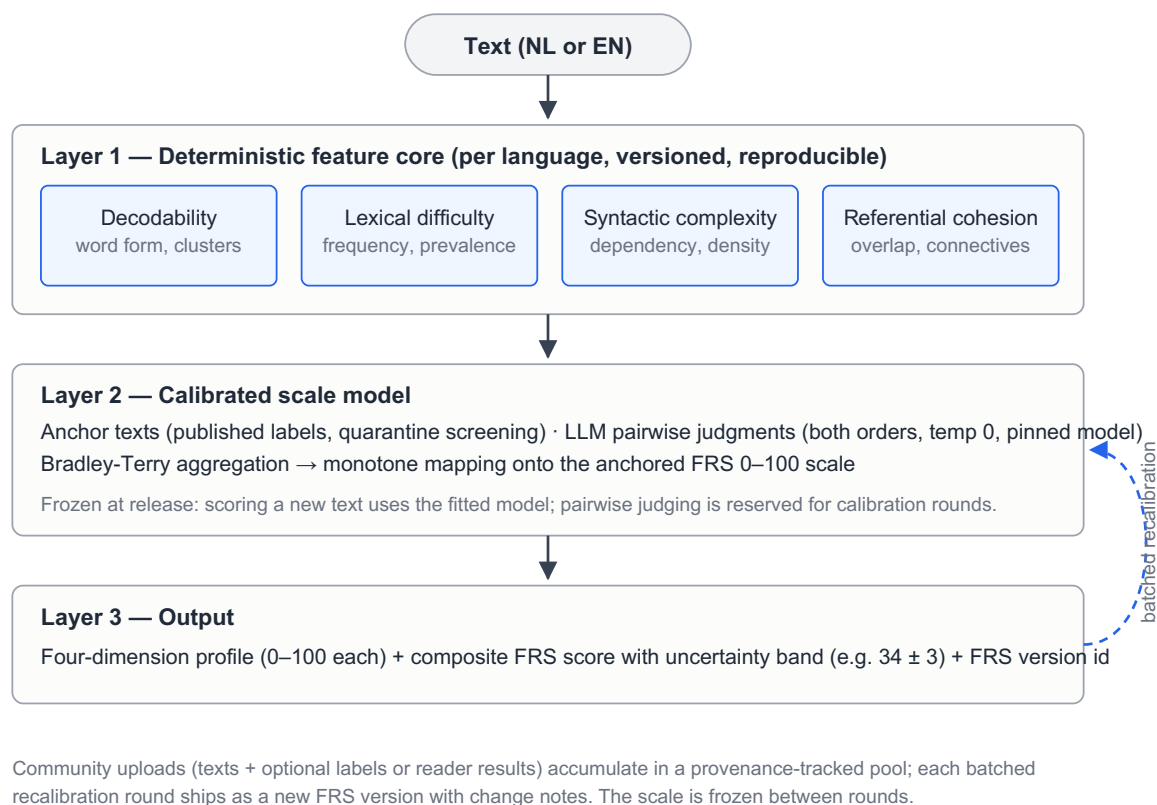


Figure 2. The FRS architecture: deterministic feature core, anchored pairwise calibration, and versioned output, with community contributions feeding batched recalibration rounds.

### 3.5 Openness requirements

The standard's normative core is short: an implementation may call itself FRS-conformant only if (a) its feature definitions, coefficients, prompts, anchor lists, and version history are public; (b) identical inputs yield identical outputs within a version; (c) scores are reported with version and uncertainty. The reference implementation, including the harness that produced §4, is published alongside the papers.

## 4. Preliminary evidence: two pilots

Both pilots use the same materials: the twenty Dutch seed texts of the earlier FRS system, each carrying an integer difficulty label from its original compilation (label provenance discussed in §4.3), and the same judge model (gpt-4o-mini via Azure OpenAI). They are pilots in the strict

sense: they motivated design decisions and are reported with full materials in Appendix C; they do not validate the standard.

#### 4.1 Pilot 1: absolute scoring is reliable but compressed

Each text was scored ten times under the production rubric prompt (Appendix B.1) at temperature 0.2 and again at temperature 0. Repeat reliability was high: mean within-text standard deviation 0.106 (0.038 at temperature 0), ICC(1,1) = .967 (.991), split-half reliability above .99. Ordering against the reference labels was strong (Spearman  $\rho = .930$ ) but 24 of 190 text pairs were ordered inconsistently with their labels. The diagnostic finding is where those violations came from: the twenty texts' mean scores span only 2.00 to 4.57 of the nominal 0–12 range. Texts labeled hardest scored barely above mid-range texts. The model, asked for absolute placement, uses a fraction of the scale and flattens the top. This is a calibration failure, not a noise failure, and no amount of repeat-averaging fixes it.

#### 4.2 Pilot 2: pairwise judgment halves ordering errors

All 190 pairs were judged in both presentation orders (380 calls, temperature 0) under the pairwise prompt of Appendix B.2, and aggregated with a Bradley-Terry fit. The resulting ordering correlated with the reference labels at  $\rho = .944$  with 13 of 190 pairs violating label order, against .930 and 24 violations for absolute scoring on identical texts and the identical model. Presentation-order agreement was 83.7%: roughly one pair in six flipped its verdict with A/B order, which the both-orders design absorbs but a single-presentation design would not. Figure 3 shows both pilots side by side. These are the numbers behind two normative choices of §3.4: pairwise-not-absolute, and both-orders-always.



Same 20 Dutch texts, same judge model (gpt-4o-mini, temperature 0). Panel A: means of ten absolute scorings per text span only 2.00–4.57 of the nominal 0–12 rubric range (shaded band). Panel B: Bradley-Terry ranks from all 190 pairs judged in both presentation orders; ordering violations drop from 24 to 13 of 190 pairs.

*Figure 3. The two pilots on the same 20 texts and judge model. Panel A: absolute scoring uses only 2.00–4.57 of the legacy rubric's nominal range. Panel B: Bradley-Terry ranks from both-order pairwise judgments track the reference labels, with the suspected label error highlighted.*

### 4.3 What the pilots do and do not establish

They establish, on one small sample, that the known weaknesses of absolute LLM scoring (compression, coarse mid-scale discrimination) appear exactly as the literature predicts, and that the pairwise remedy the literature suggests does improve ordering. They do not establish scale validity:  $n = 20$ , one model, one language, and reference labels of uncertain provenance. On that last point, one text labeled at difficulty 8 won every one of its 38 pairwise comparisons, including against all texts labeled 10, reproducibly. Either the model misjudges that text consistently, or the label is wrong. We flag it rather than resolve it, and draw the general lesson stated in §3.4: anchor label quality must be treated as part of the standard, with provenance published.

---

## 5. Discussion

**Relation to Lexile and CLIB.** FRS adopts the central lesson of both conjoint scales (a formula becomes a scale through anchoring) while differing on what it asks of the world. Lexile and the CLIB test infrastructure earned their scale semantics through large-scale reader response data, which FRS does not yet have; FRS earns provisional semantics through anchors carrying trusted labels, and states the reader-side program as future work. The trade is openness: both incumbent conjoint scales are proprietary at the point of use, and FRS's entire parameter surface is published. A planned path exists to close the reader-side gap: classroom fluency measurements (accuracy, words correct per minute) collected with consent through reading tools that FRS integrates with would provide exactly the readers-by-texts response data that conjoint calibration requires. That program, including its consent and privacy architecture, is deliberately out of scope here.

**The LLM judge.** Substituting an LLM for a thousand human raters is the step that makes open calibration affordable, and it is also the step most in need of scrutiny. Whose difficulty judgment does a language model encode? The Dutch expert-judgment literature warns that even human experts disagree and project their own reading experience onto texts (a critique developed by Lentz in the Dutch readability-tools debate); an LLM judge inherits some unknown mixture of such judgments from its training data. Three design features contain the risk without eliminating it: the judge is confined to relative judgments, whose aggregation is invariant to monotone drift in the judge's internal scale; the scale is pinned to anchors whose labels do not come from the judge; and validation against human-derived gold standards (the CLEAR scores, and planned Dutch human ratings) is a stated precondition for any FRS re-

lease, in the companion paper. If the LLM judge cannot track human-derived difficulty orderings, the method fails honestly and visibly at that gate.

**Cross-language semantics.** FRS ships as two parallel scales with identical construction and per-language anchors. The temptation to declare them one scale is resisted deliberately: the norming literature shows that even within one language, reader norms fail to cross a national border (Leysen et al., 2018), and no data yet exists to link Dutch and English text difficulty empirically. A linked scale is a research goal, not a launch feature.

**License boundaries.** The English calibration path uses CLEAR, which is CC BY-NC-SA licensed. FRS's validation studies fall comfortably within non-commercial research use; any commercial deployment of CLEAR-derived calibration artifacts would not, and the specification therefore requires that released production calibrations be traceable to license-compatible anchor data. This constraint is stated openly because license laundering through model weights is a live problem in the field, and an openness standard should not commit it.

---

## 6. Limitations

Stated plainly. (1) No reader-side validation exists yet; until the companion validation paper delivers, FRS scores are calibrated predictions, not empirically linked to reader outcomes. (2) The pilots are small (20 texts, one model, Dutch only) and their reference labels have uncertain provenance, including one probable label error. (3) The anchor supply for Dutch is thin and partly license-encumbered; the standard's Dutch low end currently leans on a discrete system (AVI) whose own text-side index uses only two variables. (4) LLM judges drift across model versions; the versioning regime contains this administratively, not scientifically, and longitudinal judge-stability data does not yet exist. (5) The construct is narrower than reading: technical difficulty only, and even within that construct, text features explain a minority of reader outcome variance in the best available Dutch evidence (Kleijn, 2018). (6) Several supporting citations in this draft rest on secondary sources, marked as such in the bibliography, and are being upgraded to primary verification before submission.

---

## 7. Conclusion and future work

This paper specified the Free Reading Standard: an open, continuous, versioned measurement of technical reading difficulty for Dutch and English texts, reported as an explainable four-dimension profile plus an anchored composite with uncertainty, calibrated through pairwise LLM judgments aggregated with Bradley-Terry against published anchors. The specification is complete enough to implement (Appendices A–D) and honest about what it is not: validated. The companion paper takes up validation at scale, with three registered questions: whether LLM pairwise judgments reproduce the human-derived CLEAR difficulty ordering on English;

whether the same protocol transfers to Dutch anchors; and whether the released scale model predicts held-out anchor placements within its stated uncertainty. Two further programs follow the validation work. First, a contribution loop: the public FRS service will accept uploaded texts with optional metadata (a contributor's own difficulty judgment, observed reader results, audience information), accumulating a provenance-tracked pool that feeds periodic batched recalibration rounds, each released as a new FRS version with published change notes; the scale stays frozen between rounds, preserving reproducibility while letting the standard learn from use. Second, the conjoint ambition: consented classroom fluency data linking readers and texts on one scale, which would make FRS for Dutch what no open instrument has been in any language.

---

## References

- School Renaissance Institute. (2000). *The ATOS readability formula for books and how it compares to other formulas* (ED449468). ERIC. <https://eric.ed.gov/?id=ED449468>
- Brouwer, R. H. M. (1963). [Leesindex A]. *Pedagogische Studiën*, 40, 454–464. [Original per-word formula; rescaled per-100-words in the 1974 extension, *Pedagogische Studiën*, 51, 545–557. Article title pending verification against the journal.]
- Brysbaert, M., Mander, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51, 467–479. <https://doi.org/10.3758/s13428-018-1077-9>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911.
- 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. <https://doi.org/10.3758/s13428-022-01802-x>
- Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Wetenschappelijke verantwoording Begrijpend lezen voor groep 3 tot en met 6* (Cito LOVS). Cito.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233. <https://doi.org/10.1037/h0057532>
- Hiebert, E. H. (2024). Enhancing opportunities for decoding and knowledge building through beginning texts. *The Reading Teacher*, 77(6), 965–974. <https://doi.org/10.1002/trtr.2303> [Content cited via secondary summary; primary text paywalled at time of writing.]
- Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650.

- Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8), 1665–1692. <https://doi.org/10.1080/17470218.2015.1022560>
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Naval Air Station Memphis.
- Kleijn, S. (2018). *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing* [Doctoral dissertation, Utrecht University]. LOT. ISBN 978-94-6093-277-9.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. <https://doi.org/10.3758/s13428-012-0210-4>
- Lee, B. W., Jang, Y. S., & Lee, J. H.-J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *Proceedings of EMNLP 2021*. arXiv:2109.12258.
- Lee, B. W., & Lee, J. H.-J. (2023). Prompt-based learning for text readability assessment. *Findings of EACL 2023*. arXiv:2302.13139.
- 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.
- National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts, Appendix A: Research supporting key elements of the standards*.
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94(2), 143–174. <https://doi.org/10.1348/000712603321661859>
- Sheehan, K. M., Flor, M., Napolitano, D., & Ramineni, C. (2015). Using TextEvaluator to quantify sources of linguistic complexity in textbooks targeted at first-grade readers over the past half century (RR-15-38). *ETS Research Report Series*. <https://doi.org/10.1002/ets2.12085>
- Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: De ontwikkeling van een domeingericht meetinstrument* [Doctoral dissertation, Universiteit Twente]. Cito. ISBN 90-801795-2-3.
- Staphorsius, G., Krom, R., & De Geus, W. (1988). *Frequenties van woordvormen en letterposities in jeugddlectuur*. Cito.
- Stenner, A. J. (1996, February). *Measuring reading comprehension with the Lexile Framework*. Paper presented at the California Comparability Symposium. (ERIC ED435977)
- Stenner, A. J., Burdick, H., Sanford-Moore, E. E., & Burdick, D. S. (2007). *The Lexile Framework for Reading technical report*. MetaMetrics.
- Trokhymovych, M., Sen, I., & Gerlach, M. (2024). An open multilingual system for scoring readability of Wikipedia. *Proceedings of ACL 2024*. arXiv:2406.01835.
- Trott, S., & Rivière, C. (2024). *Measuring and modifying the readability of English texts with GPT-4*. arXiv:2410.14028.

Vajjala, S. (2021). *Trends, limitations and open challenges in automatic readability assessment research*. arXiv:2105.00973.

Vajjala, S., & Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. *Proceedings of the 13th BEA Workshop*. Association for Computational Linguistics.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3–29. <https://doi.org/10.1037/0033-2909.131.1.3>

*Bracketed notes mark bibliographic details not yet verified against the primary source; they are resolved before any published version.*

## Appendix A. Feature definitions and resources (v0.1 draft)

Per-language operational definitions of the four dimensions' features, the lexical resources they draw on (with versions and licenses), and normalization procedures. *Status: feature list fixed per §3.2; operational parameter values to be frozen with the reference implementation and published in full in v0.2.*

| Resource                                            | Language | Role                           | License/access                             |
|-----------------------------------------------------|----------|--------------------------------|--------------------------------------------|
| SUBTLEX-NL (Keuleers et al., 2010)                  | NL       | word frequency                 | free for research                          |
| Dutch word prevalence (Keuleers et al., 2015)       | NL       | lexical familiarity            | free for research                          |
| BasiLex lexicon                                     | NL       | per-grade frequency            | registration (lvdNT); license under review |
| SUBTLEX-US/UK                                       | EN       | word frequency                 | freely redistributable                     |
| Prevalence, AoA, concreteness norms (Brysbaert lab) | EN       | lexical familiarity            | free for research                          |
| CLEAR corpus (Crossley et al., 2023)                | EN       | calibration/validation anchors | CC BY-NC-SA 4.0 (non-commercial)           |

## Appendix B. Judgment prompts (verbatim)

### B.1 Historical absolute rubric (FRS v2 production prompt, baseline condition of §4.1; Dutch)

Bepaal het FRS-leesnivea (0.00–12.00) van de aangeleverde Nederlandse tekst. Gebruik onderstaande, cumulatieve indicatoren. Kies het dichtstbijzijnde niveau op stappen van 0.02; afronden naar twee decimalen.

- 0.00–0.20 = losse letters / klankzuivere éénlettergrepen (geen zinnen)
- 0.20–0.40 = groepjes van 2–4 woorden; max. 3-woorden-zinnen
- 0.40–0.80 = mini-zinnen 3–6 woorden; uitsluitend éénlettergrep of ► 1% twee-lettergrep
- 0.80–1.20 = zinsgem. 5–8 woorden; simpele één- of tweelettergrepen
- 1.20–2.00 = 6–12 woorden/zin; 2-lettergrepen en enkele samengestelde woorden
- 2.00–3.00 = 8–14 woorden/zin; mix frequente + middel-moeilijke 3-lettergrepen
- 3.00–4.00 = 10–18 woorden/zin; volle basiswoordenschat; soms vaktaal
- 4.00–6.00 = 14–24 woorden/zin; vak-/abstracte terminologie; meerdere alinea's
- 6.00–8.00 = 16–28 woorden/zin; idiomen, stilistische variatie, complexe structuur
- 8.00–10.00 = 18–35 woorden/zin; specialistische of academische taal
- 10.00–12.00 = 20–40+ woorden/zin; zeer dense of oudere/archaische teksten

Let op:

- baseer je oordeel op zinslengte, woordlengte, vak/abstracte terminologie, tekststructuur (alinea's, bijzinnen) én impliciete conceptdichtheid.
- gebruik indien nodig 1 decimaal extra om subtiele verschillen vast te leggen (bv. 3.42 is toegestaan).
- antwoord **\*\*uitsluitend\*\*** het eindcijfer (bv. `2.60`). Geen tekst, uitleg of tekens eromheen.

(English gloss: determine the FRS reading level 0.00–12.00 of the supplied Dutch text using cumulative indicators from single decodable words through dense archaic prose; answer only the final number.)

### B.2 Pairwise judgment prompt (pilot 2, §4.2; Dutch)

System message:

Je vergelijkt twee Nederlandse teksten op TECHNISCHE leesbaarheid (decodeerbaarheid en leesvloeiendheid, niet inhoud). Antwoord uitsluitend 'A' of 'B': de moeilijkste tekst.

User message: the two texts labeled Tekst A and Tekst B (truncated to 150 words each), followed by: Welke tekst is moeilijker? Antwoord alleen 'A' of 'B':. Temperature 0, maximum 3 output tokens, every pair presented in both A/B orders.

(English gloss: you compare two Dutch texts on technical reading difficulty, decodability and fluency, not content; answer only 'A' or 'B', the harder text.)

## Appendix C. Pilot materials and results

Materials: 20 Dutch texts with integer reference labels from the FRS v2 seed library. Judge: gpt-4o-mini (Azure OpenAI, API version 2025-01-01-preview). Pilot 1: 20 texts  $\times$  10 repetitions  $\times$  2 temperature conditions = 400 calls; Pilot 2: 190 pairs  $\times$  2 orders = 380 calls; zero API or parse errors. Full per-text tables, raw judgment logs, and analysis code are published with the reference implementation ( `experiments/scoring_baseline/`: `results.json`, `pairwise_results.json`, `REPORT.md` ). Headline numbers appear in §4; the per-text result tables will be reproduced in this appendix in v0.2.

---

## Appendix D. Anchor set construction (procedure)

1. Candidate collection with provenance record (source, label authority, license).
  2. Label verification: anchors enter only with a documented label source; anchors failing consistency screening (systematic disagreement with surrounding anchors under pairwise judgment, as in the §4.3 case) are quarantined for human review rather than silently retained or dropped.
  3. Stratification across the target range per language.
  4. Publication: the anchor list, labels, provenance, and every quarantine decision ship with each FRS version.
- 

## Appendix E. Glossary

AVI (Dutch discrete technical-reading level system, twelve levels); CILT (Cito's two-variable text index for technical reading); CLIB (Cito's readability index and conjoint comprehension scale); DMT (Drie-Minuten-Toets, timed word-reading test); WCPM (words correct per minute); Bradley-Terry model (paired-comparison model yielding latent strength values); OPLM (One-Parameter Logistic Model, the IRT model used in Cito's calibrations); conjoint scale (one scale expressing both reader ability and text difficulty).

---

*Change notes v0.2: figures 1–4 added; scale specification updated from the retired 0–12 prototype scale to FRS 0–100 full-spectrum with the Dutch AVI mapping table; contribution loop added to future work; verification-pass citation fixes applied.*