Technical Appendix: Metrics, Data Structures, and Estimation Procedures

Matt Krasnow, Will Krasnow

August 19, 2025

Contents

Abstract

This appendix formally defines the data structures, variables, and statistical procedures used to compute all reported metrics. It covers (i) the relational record and JSON evaluation payload, (ii) derived variables required by the estimation procedures, and (iii) the exact formulas for all sub-metrics and composites: Teacher Professional Growth (TPG), Principal Performance and Efficiency (PPE), and Relational Trust (RT). The document is written for academic advisors and statisticians and aims to be implementation-ready.

1 Notation

- Indices: teachers $i \in \{1, \ldots, N_T\}$, evaluators/raters $j \in \{1, \ldots, N_R\}$, components $k \in \mathcal{K}$, domains $d \in \mathcal{D}$, artifacts (evaluations) $a \in \{1, \ldots, N_A\}$, time periods t.
- Scores: component score $X_{ijka} \in \mathbb{R}$ (ordinal/polytomous on a bounded scale); domain score S_{ida}; recomputed overall score S_{ia}^{overall}.
- Rasch/MFRM parameters: teacher ability θ_i, rater severity ρ_j, step thresholds β_{kx} (category x for component k), optional period effect τ_t.
- Timestamps: observation/creation T_a^{obs}, draft start T_a^{draft}, finalize T_a^{final}, delivery/share T_a^{deliv}.
- Text: component summaries Y_{ka}, overall summary Y_a^{overall}, low-inference notes N_a.
- Operators: $\text{Var}(\cdot)$, $\text{SE}(\cdot)$, $\Phi(\cdot)$ standard normal CDF.

2 Data Structures

2.1 Relational Record (Evaluation Artifact)

Each evaluation artifact is a row with the following fields (logical names shown; actual column names may differ):

Field	Type	Definition / Use
`id`	`bigint`	Unique artifact identifier.
`created_at`	`timestamptz`	Creation time; default proxy for observation time T_a^{obs}.
`updated_at`	`timestamptz`	Last update time; used in latency derivations.
`shared_at`	`timestamptz` (nullable)	Delivery/share time; preferred for T_a^{deliv} if set.
`evaluation_id`	`uuid`	Logical evaluation/observation UID.
`evaluator`	`uuid`	Evaluator/rater identifier j.
`teacher_id`	`uuid` (nullable)	Teacher identifier i.
`teacher_name`	`text` (nullable)	Teacher label (for human-readable reports).
`school_id`	`uuid` (nullable)	School identifier.
`school_name`	`text` (nullable)	School label.
`organization_id`	`uuid` (nullable)	Organization/district identifier.
`framework_id`	`text`	Instructional framework key.
`evaluation`	`json`	JSON payload (Section 2.2).
`ai_evaluation`	`jsonb` (nullable)	Optional AI generation/telemetry.
`metadata`	`jsonb` (nullable)	Optional workflow metadata (timestamps, flags).
`low_inference_notes`	`text` (nullable)	Raw evidence notes N_a.
`additional_comments`	`jsonb` (nullable)	Auxiliary comments (e.g., goals).
`is_shared`	`boolean`	Whether artifact was shared.
`deleted_at`	`timestamptz` (nullable)	Soft delete timestamp (exclude if set).

2.2 Evaluation JSON Payload

The JSON conforms to the following schema:

```
Evaluation {
  domains: { [domain_id: string]: DomainEvaluation },
  metadata: EvaluationMetadata,
  summary?: string,
  summaryScores: SummaryScores
}

DomainEvaluation {
  name: string,
  components: { [component_id: string]: ComponentEvaluation },
  weight: number,
```

```
  isManuallyScored: boolean,
  summary?: string,
  domainScore: number
}

ComponentEvaluation {
  score: number,
  summary: string,
  error?: string,
  isManuallyScored: boolean,
  modified?: boolean,
  insufficientEvidence?: boolean,
  teacherSummary?: string
}

SummaryScores {
  overallScore: number,
  domainWeights: { [domain_id: string]: number }
}

EvaluationMetadata {
  framework_id: string,
  framework_name: string
}
```

2.3 Optional Metadata Keys (if present)

- `metadata.observed_at` (timestamptz): preferred T_a^{obs}.
- `metadata.draft_started_at`, `metadata.draft_finalized_at` (timestamptz): authoring interval.
- `metadata.delivered_to_teacher_at` (timestamptz): delivery time.
- `metadata.is_ai_assisted` (bool): AI assist flag.
- `metadata.source_note_ids[]` (array): linked evidence anchors.
- `metadata.prior_goal_ids[]` (array): referenced coaching goals.
- `metadata.token_count`, `metadata.word_count` (int).
- `metadata.artifact_type` (text): e.g., *observation*, *coaching*.

3 Derived Variables and Defaults

$$T_a^{\text{obs}} := \begin{cases} \texttt{metadata.observed\_at} & \text{if present} \\ \texttt{created\_at} & \text{otherwise} \end{cases}$$

$$T_a^{\text{deliv}} := \begin{cases} \texttt{metadata.delivered\_to\_teacher\_at} & \text{if present} \\ \texttt{shared\_at} & \text{else if present} \\ \texttt{updated\_at} & \text{fallback} \end{cases}$$

Word and token counts combine all component summaries and the optional overall summary:

$$\text{Words}_a := \text{wc}(Y_a^{\text{overall}} \,\|\, \{Y_{ka}\}_k), \qquad \text{Tokens}_a := \text{tc}(Y_a^{\text{overall}} \,\|\, \{Y_{ka}\}_k).$$

4 Scoring and Re-aggregation

Let w_d be the domain weight; S_{ida} the domain score for teacher i in artifact a. The recomputed overall score is

$$S_{ia}^{\text{overall}} = \sum_{d \in \mathcal{D}} w_d \, S_{ida}, \quad \text{where} \quad w_d = \begin{cases} \texttt{evaluation.summaryScores.domainWeights[d]} & \text{if provided} \\ \texttt{evaluation.domains[d].weight} & \text{fallback} \end{cases} \tag{1}$$

5 Many-Facet Rasch Measurement (MFRM)

We use a partial-credit multi-facet model to estimate teacher ability (θ_i), rater severity (ρ_j), and step thresholds (β_{kx}). For component k with ordered categories $x = 0, \ldots, m_k - 1$, define

$$\text{logit}\, \mathbb{P}(X_{ijka} \geq x) = \theta_i - \rho_j - \beta_{kx} - \tau_{t(a)}, \tag{2}$$

where τ_t is an optional period effect for time $t(a)$ determined by T_a^{obs}. Estimation proceeds via joint/conditional maximum likelihood or marginal ML with appropriate identifiability constraints (e.g., sum-to-zero over facets).

5.1 Teacher Growth and Uncertainty

For teacher i, define baseline $T0$ and follow-up $T1$ windows. With person estimates $\widehat{\theta}_i(T0)$, $\widehat{\theta}_i(T1)$ and standard errors $\text{SE}[\widehat{\theta}_i(T0)]$, $\text{SE}[\widehat{\theta}_i(T1)]$:

$$g_i = \widehat{\theta}_i(T1) - \widehat{\theta}_i(T0), \tag{3}$$

$$\text{SE}(g_i) = \sqrt{\text{SE}^2[\widehat{\theta}_i(T1)] + \text{SE}^2[\widehat{\theta}_i(T0)]}, \tag{4}$$

$$\text{CI}_{95\%}(g_i) = g_i \pm 1.96 \, \text{SE}(g_i). \tag{5}$$

Group growth aggregates via inverse-variance weighting:

$$\bar{g} = \frac{\sum_i w_i g_i}{\sum_i w_i}, \quad w_i := \text{SE}(g_i)^{-2}, \qquad \text{SE}(\bar{g}) = \left(\sum_i w_i \right)^{-1/2}. \tag{6}$$

5.2 Improvement Index (Percentile Translation)

Let SD_{pre} be the standard deviation of $\widehat{\theta}_i$ in the baseline window. Define standardized effect $d := g/SD_{\text{pre}}$ and percentile translation

$$U3 = 100 \times \Phi(d). \tag{7}$$

6 Reliability

6.1 Generalizability Theory (G-Study and D-Study)

With persons (p), raters (r), and occasions (o), random-effects ANOVA yields variance components $\sigma^2_p, \sigma^2_{pr}, \sigma^2_{po}, \sigma^2_{pro}$. For planned numbers of raters n_r and occasions n_o, the relative G-coefficient is

$$G = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pr}/n_r + \sigma^2_{po}/n_o + \sigma^2_{pro}/(n_r n_o)}. \tag{8}$$

The D-study solves for (n_r, n_o) to achieve target G (e.g., $G \geq 0.80$).

6.2 Rasch/MFRM Person Separation

From the MFRM, report person separation reliability (PSI) for θ estimates and the number of distinguishable strata:

$$\text{Strata} \;=\; \frac{4 \times \text{Separation} + 1}{3}. \tag{9}$$

7 Fairness and Drift

- **Rater severity spread:** $\text{SD}(\{\widehat{\rho}_j\}_j)$. Flag unusually wide spreads.
- **Extremes:** $\max_j |\widehat{\rho}_j|$.
- **Drift:** segment by month/quarter via T_a^{obs} and examine $\widehat{\rho}_j(t)$; apply control charts or re-fit time-sliced models.
- **Differential Rater Functioning (DRF):** test interactions of raters with subgroups (e.g., school/subject if available in metadata).

8 Text-Based Validity, Alignment, and Tone

8.1 Evidence-to-Rating Alignment Index (ERAI)

For each component k:

$$\text{cit\_presence}_{ka} \in \{0,1\} \quad \text{(regex over summaries/notes)} \tag{10}$$

$$\text{rubric\_sim}_{ka} \in [0,1] \quad \text{(embedding cosine to rubric text)} \tag{11}$$

$$\text{entail}_{ka} \in [0,1] \quad \text{(NLI: notes} \Rightarrow \text{summary claims)} \tag{12}$$

$$\text{consist}_{ka} \in [0,1] \quad \text{(text-inferred level vs. numeric score)} \tag{13}$$

Component-level alignment:

$$\text{ERAI}_{ka} = \tfrac{1}{4}(\text{cit\_presence}_{ka} + \text{rubric\_sim}_{ka} + \text{entail}_{ka} + \text{consist}_{ka}). \tag{14}$$

Artifact-level alignment averages valid components:

$$\mathrm{ERAI}_a = \frac{1}{|\mathcal{K}_a|} \sum_{k \in \mathcal{K}_a} \mathrm{ERAI}_{ka}, \quad \mathcal{K}_a := \{k : \text{not } \texttt{insufficientEvidence}\}. \tag{15}$$

8.2 Clarity of Expectations

Within artifact text (overall + components), compute

$$\mathrm{clarity}_a = \mathbf{1}\{\text{time-bound}\} + \mathbf{1}\{\text{measurable}\} + \mathbf{1}\{\text{rubric-linked}\} \in \{0, 1, 2, 3\}.$$

8.3 Relational Tone Signals

On concatenated text:

$$\mathrm{affirm\_share}_a = \frac{\#\text{affirming tokens}}{\mathrm{Tokens}_a}, \tag{16}$$

$$\mathrm{praise:suggest}_a = \frac{\#\text{praise sentences}}{\#\text{suggestion sentences}}, \tag{17}$$

$$\mathrm{we/you}_a = \frac{\#\text{“we”}}{\#\text{“you”}}, \tag{18}$$

$$\mathrm{hedge\_rate}_a = 1000 \times \frac{\#\text{hedges}}{\mathrm{Tokens}_a}, \tag{19}$$

$$\mathrm{follow\_through}_a \in \{0, 1\} \quad \text{(references to prior goals)}. \tag{20}$$

8.4 Topic Alignment (School-Level)

Let $\boldsymbol{\pi}^{(\mathrm{FG})}$ be the topic distribution from teacher focus groups and $\boldsymbol{\pi}^{(\mathrm{FB})}$ from principal feedback text. Define

$$\mathrm{topic\_align} = \frac{\boldsymbol{\pi}^{(\mathrm{FG})} \cdot \boldsymbol{\pi}^{(\mathrm{FB})}}{\|\boldsymbol{\pi}^{(\mathrm{FG})}\|_2 \, \|\boldsymbol{\pi}^{(\mathrm{FB})}\|_2} \in [0, 1]. \tag{21}$$

9 Time, Throughput, and Pipeline

9.1 Authoring Time and Latency

$$\mathrm{authoring\_minutes}_a = \begin{cases} (T_a^{\mathrm{final}} - T_a^{\mathrm{draft}})/60 & \text{if both present} \\ \mathrm{N/A} & \text{otherwise} \end{cases} \tag{22}$$

$$\mathrm{latency\_days}_a = (T_a^{\mathrm{deliv}} - T_a^{\mathrm{obs}})/86400. \tag{23}$$

9.2 Throughput, Coverage, Streak

Let month $m(a) := \mathrm{date\_trunc}(\mathrm{month}, T_a^{\mathrm{deliv}})$ (or T_a^{obs} if delivery not used).

$$\mathrm{throughput}_{jm} = \#\{a : \text{evaluator } j,\, m(a) = m\}, \tag{24}$$

$$\mathrm{coverage}_{sm} = \frac{\#\{\text{distinct } i \text{ at school } s \text{ with } \geq 1 \text{ artifact in } m\}}{\text{teacher roster size at } s \text{ in } m}, \tag{25}$$

$$\mathrm{streak}_i = \max \text{ run length of consecutive months with } \geq 1 \text{ artifact for } i. \tag{26}$$

9.3 Pipeline Efficiency

$$\text{converted\_7d\_rate} = \frac{\#\{a : T_a^{\text{deliv}} - T_a^{\text{obs}} \le 7 \text{ days}\}}{\#\{a\}}, \tag{27}$$

$$\text{edits\_to\_final}_a \approx \sum_{k \in \mathcal{K}_a} \mathbf{1}\{\texttt{modified} = \text{true}\} \quad \text{(proxy if version logs absent)}. \tag{28}$$

10 Coaching and Goals

$$\text{coaching\_per\_teacher\_term} = \frac{\#\{\text{coaching artifacts for teacher in term}\}}{\text{term length}}, \tag{29}$$

$$\text{goals\_set\_per\_teacher} = \#\{\text{new goal identifiers in term}\}, \tag{30}$$

$$\text{SMART\_score} \in \{0, \dots, 5\} \text{ (presence of S/M/A/R/T attributes via rules/classifier)}. \tag{31}$$

11 AI-Assist Utilization and Effect (if logged)

$$\text{utilization\_rate} = \mathbb{E}[\mathbf{1}\{\texttt{is\_ai\_assisted}\}], \tag{32}$$

$$\Delta\text{time} = \text{median}(\text{authoring\_minutes} \mid \text{assisted}) - \text{median}(\text{authoring\_minutes} \mid \text{matched non-assisted}), \tag{33}$$

where matching controls for *text length* (words/tokens) and *component coverage* within evaluator j. Report Hodges–Lehmann estimate and Wilcoxon p-value.

12 Focus Groups and Surveys

12.1 Trust Index

Code focus-group transcripts for *respect*, *competence*, *integrity*, and *personal regard*. With intensity $c \in \{0, 1, 2\}$ for each coded occurrence and $C_{\max}$ the maximum possible per participant:

$$\text{RT}_{\text{FG}} = \frac{\sum \text{code intensities}}{C_{\max} \times \#\text{participants}}. \tag{34}$$

Report teacher and principal indices separately and their alignment gap.

12.2 Usability and Workload

Compute standard SUS (0–100) on the 10-item instrument and optional NASA-TLX workload composites for the authoring task.

13 Comparative Designs

13.1 Matched Rollout & Difference-in-Differences

For outcome Y_{st} (e.g., school-level mean $\widehat{\theta}$ or latency), let D_{st} indicate adoption. Estimate

$$Y_{st} = \alpha_s + \gamma_t + \delta \cdot D_{st} + \boldsymbol{X}_{st}^{\top}\boldsymbol{\beta} + \varepsilon_{st}, \tag{35}$$

with school fixed effects α_s, period fixed effects γ_t, robust SEs clustered by school. For staggered adoption, use group-time average treatment effects (e.g., Callaway–Sant'Anna estimators) and event-study pre-trend checks.

13.2 Stepped-Wedge (Non-Random or Randomized)

Fit a mixed model with period fixed effects and cluster (school) random intercepts; report ITT effects, intra-class correlation (ICC), and pertinent sensitivity analyses.

14 Composite Indices as 0-Based Difference Scores

Composites are defined as weighted sums of *natural-unit* differences from a pre-specified baseline, centered at 0 so that positive values indicate improvement and negatives indicate decline. No z-score standardization or percentage rescaling is used.

14.1 Baselines and Orientation

Let b index the baseline window (e.g., pre-period) or control condition, and t the follow-up window for a given reporting unit (teacher, evaluator, or school as appropriate). For each sub-metric M we compute

$$\Delta M := \bar{M}_t - \bar{M}_b, \tag{36}$$

with direction aligned so that higher is better. For "lower-is-better" metrics (e.g., authoring minutes, latency), use $\Delta M := -(\bar{M}_t - \bar{M}_b)$.

Unit scales. To place heterogeneous natural units on a comparable contribution scale while preserving interpretability, each sub-metric has a fixed, pre-specified *unit scale* u_M (not estimated from the data). The scaled difference is $\Delta M/u_M$, where one unit corresponds to a meaningful change (e.g., $u_{\text{authoring}} = 10$ minutes, $u_{\text{latency}} = 1$ day, $u_{\text{throughput}} = 1$ artifact/month, $u_{\text{ERAI}} = 0.05$ absolute, $u_g = 0.10$ logits). Chosen u_M values should be pre-registered and held constant across reports.

14.2 TPG — Teacher Professional Growth

Primary construct is growth in rater-adjusted latent performance, supplemented by validity and reliability guardrails. Define

$$\Delta g := \bar{g}_t - \bar{g}_b \quad \text{(logits; already oriented)}, \tag{37}$$

$$\Delta\text{ERAI} := \overline{\text{ERAI}}_t - \overline{\text{ERAI}}_b, \tag{38}$$

$$\Delta\text{Reliability} := \overline{\text{PSI}}_t - \overline{\text{PSI}}_b, \tag{39}$$

$$\Delta\text{Fairness} := -\Big(\text{SD}(\{\widehat{\rho}_j\})_t - \text{SD}(\{\widehat{\rho}_j\})_b\Big), \tag{40}$$

$$\Delta\text{FG growth} := \overline{\text{FG growth index}}_t - \overline{\text{FG growth index}}_b. \tag{41}$$

With weights $\boldsymbol{w}^{(\text{TPG})}$ (summing to 1) and unit scales u_M, the composite is

$$\text{TPG} = \sum_{M \in \{g,\text{ERAI},\text{Reliability},\text{Fairness},\text{FG growth}\}} w_M^{(\text{TPG})} \cdot \frac{\Delta M}{u_M}. \tag{42}$$

Recommended default emphasizes growth: $w_g^{(\mathrm{TPG})} = 0.6$, others share the remainder unless pre-registered otherwise.

14.3 PPE — Principal Performance and Efficiency

Define oriented differences for key efficiency and production metrics:

$$\Delta\text{authoring} := -\big(\overline{\text{authoring\_minutes}}_t - \overline{\text{authoring\_minutes}}_b\big), \tag{43}$$

$$\Delta\text{latency} := -\big(\overline{\text{latency\_days}}_t - \overline{\text{latency\_days}}_b\big), \tag{44}$$

$$\Delta\text{throughput} := \overline{\text{artifacts/leader/month}}_t - \overline{\text{artifacts/leader/month}}_b, \tag{45}$$

$$\Delta\text{coverage} := \overline{\%\,\text{teachers with} \geq 1\ \text{artifact/month}}_t - \overline{\%\,\text{teachers with} \geq 1\ \text{artifact/month}}_b, \tag{46}$$

$$\Delta\text{content} := \overline{\text{content richness index}}_t - \overline{\text{content richness index}}_b, \tag{47}$$

$$\Delta\text{pipeline} := \overline{\text{converted\_7d\_rate}}_t - \overline{\text{converted\_7d\_rate}}_b. \tag{48}$$

Composite:

$$\text{PPE} = \sum_{M \in \{\text{authoring,latency,throughput,coverage,content,pipeline}\}} w_M^{(\mathrm{PPE})} \cdot \frac{\Delta M}{u_M}, \quad \sum_M w_M^{(\mathrm{PPE})} = 1. \tag{49}$$

Recommended default emphasizes time saved and latency: $w_{\text{authoring}}^{(\mathrm{PPE})} = 0.35$, $w_{\text{latency}}^{(\mathrm{PPE})} = 0.25$, remaining weight distributed over production/quality metrics.

14.4 RT — Relational Trust

Define oriented differences:

$$\Delta\mathrm{RT}_{\mathrm{FG}} := \overline{\mathrm{RT}_{\mathrm{FG}t}} - \overline{\mathrm{RT}_{\mathrm{FG}b}}, \tag{50}$$

$$\Delta\text{Tone} := \overline{\text{Relational Tone Index}}_t - \overline{\text{Relational Tone Index}}_b, \tag{51}$$

$$\Delta\text{Clarity} := \overline{\text{clarity}}_t - \overline{\text{clarity}}_b, \tag{52}$$

$$\Delta\text{TopicAlign} := \overline{\text{topic\_align}}_t - \overline{\text{topic\_align}}_b. \tag{53}$$

Composite:

$$\text{RT} = \sum_{M \in \{\mathrm{RT}_{\mathrm{FG}},\text{Tone,Clarity,TopicAlign}\}} w_M^{(\mathrm{RT})} \cdot \frac{\Delta M}{u_M}, \quad \sum_M w_M^{(\mathrm{RT})} = 1. \tag{54}$$

Recommended default emphasizes focus-group trust: $w_{\mathrm{RT}_{\mathrm{FG}}}^{(\mathrm{RT})} = 0.5$.

14.5 Uncertainty and Reporting

Report point estimates of each ΔM in natural units alongside the composite. Compute composite confidence intervals via nonparametric bootstrap over teachers (TPG,RT) or evaluators (PPE), maintaining the weighting and unit scales within each replicate.

15 Data Quality & Missingness

- Exclude artifacts with `deleted_at` not null.
- Prefer T_a^{obs} from metadata; otherwise use `created_at`.
- If delivery is unlogged, use `shared_at` then `updated_at` for T_a^{deliv}.
- Report coverage (%) for each metric and apply pairwise deletion by default; consider multiple imputation for time-series analyses if MAR is plausible.

A Field Dictionary: Definitions, Types, and Calculations

Identity & Linkage

Name	Type	Definition / Calculation
`id`	bigint	Artifact identifier.
`evaluation_id`	uuid	Observation UID.
`teacher_id`, `teacher_name`	uuid, text	Teacher key/label (i).
`evaluator`	uuid	Evaluator/rater key (j).
`school_id`, `school_name`	uuid, text	School key/label.
`organization_id`	uuid	Organization/district key.
`framework_id`	text	Instructional framework key.

Timestamps

Name	Type	Definition / Calculation
`created_at`	timestamptz	Default T_a^{obs} if metadata absent.
`updated_at`	timestamptz	Used in latency fallback.
`shared_at`	timestamptz	Preferred T_a^{deliv} if set.
`metadata.observed_at`	timestamptz	Preferred observation time.
`metadata.draft_started_at`	timestamptz	Authoring start (if present).
`metadata.draft_finalized_at`	timestamptz	Authoring end (if present).
`metadata.delivered_to_teacher_at`	timestamptz	Delivery time (if present).

Scores & Text

Name	Type	Definition / Calculation
`domains[d].domainScore`	number	Domain score S_{ida}.
`domains[d].weight`	number	Domain weight w_d (fallback).
`summaryScores.domainWeights[d]`	number	Preferred w_d.
`components[k].score`	number	Component score X_{ijka}.
`components[k].summary`	string	Text Y_{ka}.
`summary`	string	Overall text Y_a^{overall}.

Name	Type	Definition / Calculation
summaryScores.overallScore	number	Provided overall score (audited vs. recompute).
low_inference_notes	string	Evidence text N_a.

Flags & Links

Name	Type	Definition / Calculation
components[k].isManuallyScored	bool	Manual scoring flag.
components[k].modified	bool	Edited post-generation.
components[k].insufficientEvidence	bool	Evidence insufficient.
metadata.is_ai_assisted	bool	AI assist used.
metadata.source_note_ids[]	array	Evidence anchors.
metadata.prior_goal_ids[]	array	Goal references.

Derived Quantities

Name	Definition / Formula
S_{ia}^{overall}	$\sum_d w_d S_{ida}$
Words, Tokens	Counts over $Y_a^{\text{overall}} \| \{Y_{ka}\}$
Evidence density	(#citations)/Words $\times$ 100
Authoring minutes	$(T_a^{\text{final}} - T_a^{\text{draft}})/60$ (if present)
Latency days	$(T_a^{\text{deliv}} - T_a^{\text{obs}})/86400$
Throughput	#artifacts per evaluator-month
Coverage	%teachers with ≥ 1 artifact per month
Streak	Longest consecutive months with ≥ 1 artifact

B Estimation Outputs (Summary)

Quantity	Definition / Estimation
$\widehat{\theta}_i$	Teacher ability (logits) from MFRM.
$\widehat{\rho}_j$	Rater severity (logits) from MFRM.
$\widehat{\beta}_{kx}$	Step thresholds from MFRM.
g_i, $\bar{g}$	Teacher growth and group mean; CIs via delta method.
$U3$	Percentile translation $100\Phi(g/SD_{\text{pre}})$.
G	Generalizability coefficient; D-study recommendations.
PSI, Strata	Person separation and distinguishable strata.
ERAI	Evidence-to-rating alignment index (0–1).

Quantity	Definition / Estimation
Tone/Clarity	Affirm share, praise:suggest, we/you, hedging, clarity (0–3).
Topic alignment	Cosine similarity of topic distributions (0–1).

C Implementation Notes

- All time deltas are computed in seconds and rescaled (min, d).
- When delivery time is not explicitly recorded, use `shared_at`, else `updated_at`.
- For matched comparisons (AI effect), match within evaluator on word count and component coverage.
- Bootstrap CIs: 2000 replicates by teacher for TPG/RT and by evaluator for PPE.
- Exclude artifacts with `deleted_at` not null from all analyses.