Reliability and Accuracy of a Record-Level Review Standard (JRS)

Rungs 1 and 2 of the JRS Evidence Development Program

Draft. Reproducibility (Rung 1) and reliability (Rung 2a) results are computed from collected data. Accuracy (Rung 2b) is preliminary (data collection ongoing). All numbers verified against the study database.

Abstract

Background. The Justification Review Standard (JRS) is a record-level review method that evaluates whether the documented basis for a consequential decision can be independently reconstructed from the record alone. A record that cannot be so reconstructed carries Decision Reconstruction Risk (DRR).

Objective. This paper reports the first stages of a pre-registered evidence development program for JRS, covering reproducibility (Rung 1), reliability (Rung 2a), and preliminary accuracy (Rung 2b).

Methods. Reproducibility was assessed by having three independent AI providers apply JRS to the same constructed records. Reliability was assessed among independent expert and trained-reviewer panels using Gwet's AC1 as the primary chance-corrected coefficient. Preliminary accuracy was assessed against a held-out answer key, independently verified before scoring.

Results and interpretation. Independent reviewers reached substantial agreement (Gwet's AC1 = 0.74 for experts, 0.63 for trained reviewers, across 10 records). Cross-vendor AI agreement averaged 84% across 15 constructed records. Accuracy testing remains ongoing. The findings support reproducible application and substantial inter-rater reliability; they do not establish accuracy, value, or real-world effectiveness, which are later stages.

1. Background

Records increasingly justify consequential decisions, and increasingly they are produced with AI-assisted drafting that makes them fluent without guaranteeing that the basis for a conclusion is present. JRS evaluates a record on whether its reasoning is reconstructable from the record itself, across five conditions: reconstructability, basis identification, chronology, decision-process traceability, and evidentiary sufficiency. Each record receives one of three determinations: Ready, Needs work, or Gap. JRS evaluates whether a later reviewer can reconstruct the documented basis for a decision from the record itself, without access to the author. Decision Reconstruction Risk (DRR) is the condition in which a record cannot, on its own terms, allow an independent reviewer to reconstruct the basis for a consequential decision.

2. The evidence program and this paper's scope

Validation proceeds in staged rungs, each answering one question:

```
       JRS Evidence Development Program

Rung 1   Reproducibility   do independent AI models apply JRS alike?      [84%, 15 records]
  |
Rung 2a  Reliability       do independent human reviewers agree?          [AC1 0.74 / 0.63]
  |
Rung 2b  Accuracy          do reads match a verified key?                 [preliminary]
  |
Construct validity         are the five conditions distinct dimensions?   [data ready]
  |
Rung 3   Criterion validity do flagged records fail in real cases?        [12 cases, early]
  |
```

```
External validity          does it hold on real, non-constructed records? [future]
```

This paper reports Rung 1 and Rung 2a in full, Rung 2b as preliminary, and locates the remaining rungs as future work.

3. Methods

3.1 Reproducibility (Rung 1)

Each constructed record is judged by three independent AI models, one per provider (Anthropic, OpenAI, Google), cross-vendor. The measure is how often the three independent models return the same JRS read on the same record. This runs as an automated nightly study. Cross-vendor (rather than same-provider) models are used because they do not share lineage, giving a stronger independence signal. Agreement here is explicitly not accuracy and not validation.

3.2 Reliability (Rung 2a)

Independent raters applied the five JRS conditions to a shared set of records, recording a determination (Ready / Needs work / Gap) for each. Raters worked independently. Raters whose codes begin with E are designated experts; the remainder are trained reviewers. Agreement was assessed with Gwet's AC1 as the primary chance-corrected coefficient, chosen for its robustness to the kappa paradox under skewed category marginals, with raw percent agreement reported alongside.

3.3 Accuracy (Rung 2b)

A separate 24-record set (12 constructed grounded, 12 constructed ungrounded) carries a held-out answer key. The key was fixed by an operational rule and verified by independent raters blind to the intended labels before any accuracy analysis. Reviewers judge each record blind to the key; accuracy is scored against the verified key.

3.4 Materials

Constructed records were developed to represent realistic administrative documentation while allowing known evidentiary characteristics (presence or absence of a documented basis, dated sequence, and traceable reasoning) to be systematically controlled. No real records are used at this stage.

4. Results

4.1 Reproducibility (Rung 1) — collected data

On 15 constructed records, three independent providers (Anthropic claude-opus-4-8, OpenAI gpt-5, Google Gemini) agreed on the JRS read **84%** of the time (nightly automated run, latest 2026-07-06; history ranges 78–87% as the record set expanded from 3 to 15). This is a reproducibility signal only: it shows independent models apply the read consistently, not that the read is correct.

4.2 Reliability (Rung 2a) — collected data

Across **10 records**:

Rater group	Records	Mean raters/record	Raw agreement	Gwet's AC1
Experts	10	3.6	88%	**0.74**
Trained reviewers	10	7.2	83%	**0.63**

Both coefficients fall in the "substantial" range (0.61–0.80 on the Landis–Koch scale, the convention behind the pre-registered floor of 0.61). The reviewer coefficient (0.63) clears that floor; the expert coefficient (0.74) exceeds it. Across all JRS-mode determinations (108 labels), the distribution was Gap 69%, Needs work 18%, Ready 13%,

consistent with a record set weighted toward reconstructability problems.

4.3 Accuracy (Rung 2b) — preliminary

The 24-record answer key was independently reproduced by raters blind to the intended labels, 24 of 24, fixing the key against which accuracy is scored. Reviewer completion is in progress; a full accuracy estimate (with sensitivity, specificity, and confidence intervals) is deferred until the pre-registered reviewer sample completes. Early completions are consistent with above-chance separation, reported here as preliminary and not as a confirmed result.

5. Discussion

The reliability result is the substantive finding of this stage: independent raters, expert and trained, apply JRS to records with substantial, chance-corrected agreement, and the reviewer coefficient meets the threshold set in advance. This supports the claim that JRS is applied consistently, not idiosyncratically.

Reliability is a prerequisite for later validation. Before a review method can be evaluated for accuracy or practical value, independent reviewers must first demonstrate that they can apply it consistently. The findings reported here therefore establish a methodological foundation rather than a final validation of JRS: reliability is necessary but not sufficient for the accuracy, value, and effectiveness questions that follow.

6. Limitations

- Reliability rests on 10 records; a larger set would tighten the estimate.
- Accuracy is preliminary pending reviewer completion; no confirmed accuracy claim is made here.
- Reproducibility (Rung 1) reports raw cross-vendor agreement on 15 constructed records; chance-corrected coefficients on the model votes are still to be added.
- Construct validity (whether the five conditions are distinct dimensions) is a separate analysis, addressed elsewhere.
- Records are constructed; external validity on real-world records is a later rung.
- The findings do not establish that JRS improves organizational outcomes, reduces litigation risk, or increases decision quality. Those questions require separate evaluation.

7. Pre-registered thresholds

Reliability is considered supported only if Gwet's AC1 meets 0.61 with an adequate lower bound; the reviewer result meets this, the expert result exceeds it. Accuracy and value thresholds are pre-registered and evaluated only after data lock. Failing a threshold is reported as a null or weak result and not reinterpreted.

8. Next steps

The staged program proceeds in order: complete the accuracy analysis on the 24-record set; run a standard-versus-baseline comparison to test whether JRS improves on unaided judgment; establish construct validity of the five conditions; and, finally, external validation on real (non-constructed) records with documented outcomes.

9. Attribution

The reference-panel design and the chance-corrected reliability framework (joint use of Gwet's AC1, Krippendorff's alpha, and Fleiss' kappa; treatment of the kappa paradox; pre-registered agreement floors) are methodological contributions of **Ubayet Hossain, FRM, Associate Director (Model Validation), KPMG India**. The proportionality principle referenced in the standard was surfaced by pilot reviewer **Saurabh Nanda** and is credited with permission. Statistical analyses were completed before interpretation of study findings. Responsibility for the framework as

presented rests with the author.

10. Conclusion

These findings support continued evaluation of JRS through staged validation. They should be interpreted as evidence of reproducible application and substantial inter-rater reliability, not as a final demonstration of accuracy or operational effectiveness.