

Rubrica v0.1.10: Validation Note

Evidence summary accompanying the Zenodo report on Rubrica v0.1.10

October 5, 2026 · Travis Fraser, University of California, Berkeley (Haas School of Business)

SCOPE

Rubrica grades scanned, handwritten exams against an instructor's rubric. Under the default local-Ollama configuration, stored names, SIDs, and page 0 are excluded from external grading requests; those requests contain an anonymous identifier, answer-page images, and rubric text. Because answer pages and rubrics can themselves contain incidental identifiers, instructors should review and remove that information before external grading. Production grading uses temperature 0.0 where supported, which reduces sampling randomness but does not guarantee repeatable scores. This note summarizes three studies behind the version 0.1.10 claims and states what each does and does not measure. The studies were run on development builds between March and July 2026, before the 0.1.2 pilot build, while parts of the grading prompt and its safeguards were still changing; they are carried forward from the previous version of this report and were not re-run. The grading prompt and the scoring rules are unchanged between versions 0.1.2 and 0.1.10 for a model response that is not empty and gives each question once. Version 0.1.10 changed how a question returned more than once (a label that differs only in spacing, capitals or punctuation counts as the same question) and a total returned with no per-question scores are handled, and how an instructor's score edit settles an unresolved boundary average. The results below do not measure the effect of those changes.

Three different questions. Study 1 measures concordance between Rubrica and another model under a matched-input, author-run protocol. Study 2 compares Rubrica with human-assigned reference scores. Study 3 measures variation when Rubrica grades the same exams repeatedly. None substitutes for the others.

STUDY 1: AUTHOR-RUN CROSS-MODEL COMPARISON (RUN MARCH 2026)

Design

30 exams (840 scored items) from an undergraduate economics course at UC Berkeley Haas were graded in production by Claude Sonnet 4.6 and re-graded by o4-mini (OpenAI). The models received matched rubric content and rendered answer-page images, and their outputs passed through shared Python scoring functions. The provider APIs differ in message roles, request formatting, and inference controls. The comparison implementation also uses production output to supply per-question maximum points and missing-question structure. The comparison therefore estimates observed cross-model concordance; it is not a third-party audit or an experiment in which the model is literally the only variable. The in-app comparison now defaults to Gemini 3.1 Pro. Choosing another vendor reduces, but does not remove, the risk of correlated errors.

Results

Metric	Result	Reference Threshold	Status
ICC(3,1), consistency	0.966 (95% CI 0.93 to 0.98)	descriptive bands only (Koo and Li, 2016), read with the interval	Reported
Quadratic weighted kappa, letter grades	0.883	0.70 historical e-rater benchmark; task-dependent	Reported
Within-1-point agreement	96.2% (808 of 840 items)	at least 95% internal operating criterion	Met
Exact score match	88.3% (742 of 840 items)	task-dependent	Reported
Mean absolute error	0.15 points per question	no universal threshold; scale-dependent	Reported
Mean difference	+0.64 points per exam (production minus reference)	approximate 95% CI: -0.49 to +1.76	Uncertain direction
Letter-grade agreement (raw, pre-safeguard)	70% (21 of 30 exams); every mismatch between adjacent letters	not benchmarked	See note

All nine letter-grade mismatches are between adjacent letters. Seven have a production score inside the $\pm 1.5\%$ boundary window; two fall outside it and would not be re-graded by that mechanism. Eight of the nine sit on one of the two exam versions, driven largely by one rubric item on which the models often disagreed. The safeguard runs after the snapshot this comparison measures, so post-safeguard agreement with the reference model remains unmeasured. The observed production-minus-reference sample mean difference was +0.64 points per exam. Its standard deviation was about 3.00 points, and the approximate 95% confidence interval for the mean spans -0.49 to +1.76; the direction of the population mean difference remains uncertain.

Correction. Earlier versions of this report gave ICC 0.964. That value came from an implementation that left the between-rater variance in the residual term, so a constant offset between the two models lowered a statistic meant to ignore it. Recomputed from the same data with the standard two-way formula, consistency ICC(3,1) is 0.966 (95% CI 0.93 to 0.98). The comparison ran on 9 March 2026; the report describing it was written in April 2026, and earlier versions called it the April 2026 study.

What Study 1 does not show. Agreement between two models is a structural check, not evidence of accuracy against a human grader, third-party independence, or freedom from shared error. Provider-specific request differences also prevent a literal only-the-model-changed interpretation. The full item-level report is available from the author; apart from the corrected ICC, the aggregate values above are unchanged.

STUDY 2: COMPARISON WITH HUMAN-ASSIGNED REFERENCE SCORES (JULY 2026)

Design

A macroeconomics midterm (one exam version) from another instructor, with human-assigned scores recorded in Gradescope, on an exam the author did not write. Rubrica graded the same scans; results were joined to the reference scores per question on the anonymous identifier, with a clean join for all 25 exams. The exam contained 38 questions or sub-parts totaling 46.5 points that Rubrica handled through its decomposed, per-criterion rubric, plus one 25-point essay graded holistically.

Results

ITEM SET	POINTS	MEAN ABSOLUTE ERROR VS. HUMAN	MEAN BIAS	CELLS MATCHING THE HUMAN GRADE EXACTLY
Decomposed questions (38 items)	46.5	1.26 points per exam (2.7% of the section)	−0.28 points per exam (observed sample mean)	85.4% (810 of 949 compared items)
Holistically scored essay (1 item)	25	3.88 points per exam	−3.64 points per exam (observed sample mean)	32% (8 of 25)

On this sample, differences were smaller on the decomposed section than on the single holistically scored essay. The comparison does not isolate the cause: item format, point scale, content, difficulty, position, and rubric structure all differ. The result is consistent with the hypothesis that narrower criteria help, but it is not causal evidence for decomposition. The observed essay difference was negative and largest among the strongest essays. Item-level detail is available from the author on request.

One model score (question 28 on one exam) is missing, so the structured section has 949 compared items rather than 950; missing scores are excluded from the exact-match rate (counting it as a mismatch gives 810 of 950, 85.3%). Across all items including the essay, 818 of 974 matched exactly (84.0%); earlier versions of this report gave that whole-exam figure for the structured section.

Reliability on whole-exam totals (n = 25): absolute-agreement ICC(A,1) 0.868 (95% CI 0.42 to 0.96) and consistency ICC(C,1) 0.922 (95% CI 0.83 to 0.96). The wide lower bound for absolute agreement reflects the systematic essay difference. QWK is not reported; it would need a justified mapping from these scores to letter grades.

What Study 2 does not yet show. At n = 25 the reliability intervals above are wide; their precision depends on the score distribution, expected reliability, number of raters, and sample size. This is one exam from one course, so the figures are a sample-bounded comparison, not a general error rate.

STUDY 3: TEST-RETEST STABILITY (JUNE 2026)

Design

Ten final-exam submissions were selected by deterministic stratification across exam version, batch, and score range. Each was graded three times through the production pipeline on its native rubric point scale. The study asks whether repeated runs of the same model and workflow produce similar totals and the same letter grade.

Results

METRIC	RESULT
Letter-grade stability across three runs	100% (10 of 10 exams)
Mean within-exam total-score standard deviation	0.91 points (population SD of each exam's three totals; the sample SD averages 1.11)
Highest-variance item	SA1
Earlier observed baseline	1.03 points from a separate 30-exam, three-run capture

The earlier 1.03-point value came from a different corpus and rubric, and which standard-deviation convention it used was not recorded. It provides context only and is not directly comparable to either value above.

What Study 3 does not show. This is a small test-retest study from one workspace with three repeats per exam. It does not estimate cross-course stability or accuracy against a human. Student-level results are not deposited; this report includes aggregate results to preserve privacy.

LIMITATIONS AND OPEN ITEMS

ITEM	STATUS
All three studies are author-run; no third-party replication has been completed.	Open
Study 1 measures model-against-model agreement, not accuracy.	By design
Study 2 has n = 25; its reliability intervals are wide (absolute-agreement ICC 0.42 to 0.96).	Rerun planned
Studies 1 and 2 each cover one course and exam; no cross-discipline accuracy estimate exists.	Open
Study 3 covers one workspace and three repeats per exam.	Limited scope
Figures recomputed from the original source files; corrections are noted in the text.	Checked 2026-09-22

REFERENCES

- Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21(9), 1331–1335. <https://doi.org/10.1002/sim.1108>
- Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>