The potential use of AI for conducting follow-up assessments in massive statistics courses

Eduardo León Bologna & Marcelo Vaiman

National University of Córdoba

01 octubre, 2025

Motivation

- Massive enrollment in introductory statistics course in Psychology + low number of professors: ~ 140 students per professor → difficults individualized evaluation
- Multiple-choice exams partially solve the problem, but fail to assess the ability to interpret results and draw conclusions
- Interpreting skills are evaluated asking for a written report elaborated from a provided dataset
- This leads to a large number of repetitive evaluations and variations in evaluator criteria over time due to fatigue
- Our question: can a calibrated Generative AI help evaluate written statistical work consistently and at scale?

Context

Growth of Artificial Intelligence in education

- In two directions:
  - adaptive tutoring and individualized feedback for students
  - automated assessment

Relevance for massive courses

- scalable marking
- faster feedback without sacrificing quality
- eliminate the effects of rater fatigue

Objectives

- Quantify agreement and differences between three human raters and AI (chatGPT5) when assigning marks to students written reports
- Explore task-level agreement to detect where rubrics need refinement

Course context & tasks

- Our students: Psychology undergraduates in a compulsory Statistics course
- Our assessment consists of
  - Calculate and interpret:
    - central tendency & dispersion mesures (25 pts)
    - tables and graphs (25 pts)
    - ranked frequency tables (15 pts)
    - measures of association between two variables (25 pts)
  - Present results in APA style (10 pts)

Procedure

- 3 professors graded 9 reports independently
- AI received the same reports and underwent 2-step calibration:
  1) Exemplar report (100 pts) as gold standard
  2) Targeted guidance per activity (e.g., require text, penalize missing interpretation, handle cumulative frequencies correctly, etc.)

Methods

- Build a single dataset with the four raters' scores (per task and total)
- Compute:
    - Pairwise correlations on total scores
    - Kendall's W across human raters and across all four raters
    - One-way ANOVA on total scores by rater

Results, global

High agreement overall between humans and AI

- Human–human correlations 0.74–0.79
- Human/AI correlations 0.65–0.77
- No significant mean differences in total scores across raters (t/ANOVA non-signficant)

Results, by task

Different agreement level

- Lower on the ranking task (Activity 3)
- Highest on the association task (Activity 4)

Interpretation

- When calibrated, a generative model can match expert consistency on written work
- AI can stabilize criteria and mitigate fatigue effects in large cohorts
- Task-level discrepancies flag rubric precision needs

Educational implications

- Use of AI as a complement:
  - First-pass scoring + instant formative feedback for students
  - Free up instructor time for higher-order skills: oral defense, criticisms of the choice of procedures, contextual examples
- Integrate feedback instances where students discuss AI comments with instructors

Ethic issues

- Monitor bias: continuous auditing of outputs
- Data protection: compliant workflows and minimal data exposure
- Academic integrity: distinguish using AI for evaluation from student use of AI in production

The AI–AI paradox

Paradox

- Students use AI to produce assignments
- Instructors use AI to grade them
  - Humans risk being out of the loop

Risks to assessment quality

- Loss of construct validity: the artifact is graded, not the student's understanding
- Authenticity loss: tasks optimize for model output rather than learning goals

To keep humans as protagonists

We should

- Require students to declare AI use with the submission
- Include short interviews asking to explain decisions, interpretations, and limitations, as part of the evaluation
- Ask for staged submissions with change logs to show learning progression

Assessment proposal to mitigate the “paradox”

Hybrid rubric:

- Use AI for mechanical checks: APA, presence of required elements, use of correct procedures
- Let humans evaluate reasoning, interpretation, and communication quality

Preliminary conclusion

It is feasible

- AI can deliver human-comparable marking after calibration

And useful

- speeds feedback supports in large courses

It is necessary to be prudent

- to keep humans in the loop we need at least one instance of face-to-face evaluation

Limitations & future work

Next steps

- replicate with larger samples and varied tasks
- split the evaluation sample, train on one part and test agreement human - AI on the other
- formalize calibration protocols and versioned rubrics
- study impact on learning, using face to face evaluations to compare along time

Thank you very much for your attention