

# AI Replication Games October 16th

A one hour training for quantitative social sciences

Juan Pablo Posada Aparicio

Institute for Replication, University of Ottawa

October 2025

## Session goals

- ▶ Align on the Replication Games storyline, schedule, and roles before event day.
- ▶ Clarify randomization, treatment arms, and deliverables drawn from the preregistered design.

## Roadmap, about 60 minutes

- ▶ 5 minutes. Pre-games context and storyline.
- ▶ 10 minutes. Study design, randomization, and assignments.
- ▶ 15 minutes. Event-day operations, deliverables, and support.
- ▶ 20 minutes. ChatGPT Plus feature tour across the research workflow.
- ▶ 5 minutes. Q and A.

## Pre-games storyline

- ▶ Kick off the Replication Games and align on the narrative before event day.
- ▶ Friendly human versus AI-augmented challenge tests speed, accuracy, and issue-spotting.
- ▶ We study vertical gaps across expertise tiers and horizontal gaps across disciplines.

## Team and support network

- ▶ Institute for Replication with Abel Brodeur coordinates the University of Ottawa hub.
- ▶ Support crew: Ghina Abdul Baki, Juan Pablo Aparicio, Bruno Barbarioli, Lenka Fiala, Derek Mikola, David Valenta.
- ▶ University of Ottawa hosts in person; virtual participants rely on Zoom. Email: [instituteforreplication@gmail.com](mailto:instituteforreplication@gmail.com).

# Study design essentials

- ▶ You will be randomly assigned to work either as a human-only participant or as part of the cyborg (AI-assisted) arm.
- ▶ Around 300 participants stratified by expertise tier and discipline tag.
- ▶ Randomized 1:1 within strata to AI-assisted access versus human-only control.
- ▶ Task pool spans Economics, Political Science, and Psychology with assignments balancing in- and out-of-discipline exposure.

## Treatment arms and tiers

- ▶ **Human control.** No external AI during the work window; document everything manually.
- ▶ **Cyborg arm.** ChatGPT Plus with Advanced Data Analysis, Deep Research, Agent Mode, and Codex CLI support; other AI tools are allowed if they document usage.
- ▶ Tiers from Undergraduate to Professor; we log discipline tags, coding experience, and AI familiarity for heterogeneity analyses.

# Participant prep checklist

- ▶ Complete this orientation and skim the reporting workbook (GitHub template).
- ▶ Bookmark the preregistered repository materials on OSF ([osf.io/dkfzt](https://osf.io/dkfzt)) for mirrored access.
- ▶ Accept the ChatGPT Team invite promptly.
- ▶ Confirm hardware and required software licenses (R/Stata/Python) before event day.
- ▶ Review the assignment email so you know your tier, discipline tag, arm, and team roster.



## Event-day timeline and workflow

- ▶ 8:45 local check-in or remote login; 9:00 shared Dropbox folder unlocks (OSF mirror provided for anyone without Dropbox) alongside the reporting sheet.
- ▶ Read instructions, note the focal result highlighted on the first page of the paper, and confirm you have every required file.
- ▶ Reproduce the assigned result, logging timestamps; audit code for major and minor errors.
- ▶ Run robustness checks and keep the reporting sheet—referee-report tab included—updated throughout the seven-hour window.

## Deliverables, compliance, and support

- ▶ Submit the reproduced result, error log, and reporting workbook by 16:00, plus qualitative notes if helpful.
- ▶ Control arm pledges no AI; AI arm completes a short end-of-day survey noting which AI tools they used and how often (no prompt logging required).
- ▶ Primary outcomes cover success, timing, error counts, and robustness; secondary outcomes review narratives and recommendations.
- ▶ Technical or design questions: email [instituteforreplication@gmail.com](mailto:instituteforreplication@gmail.com).

# What we mean by computational reproducibility

- ▶ Anyone with the shared code, data, and instructions should be able to rerun the workflow and obtain the same focal result.
- ▶ Re-runs should execute end-to-end without manual tweaks, with scripts producing identical figures, tables, and statistics.
- ▶ Document external dependencies (software versions, seeds, APIs) so others can recreate the original computing environment.

# Classifying coding issues

- ▶ **Major errors:** Significantly change the numerical result, invalidate inference, or change conclusions.
- ▶ **Minor errors:** Issues that do not alter the reported outcome, inference or conclusions.
- ▶ Missing file paths, hard-coded directories, or absent packages are expected, do not treat them as coding errors.

## Required robustness checks

- ▶ Each team proposes and runs two targeted robustness checks tied to the assigned result.
- ▶ Prioritize checks that stress key assumptions (e.g., alternative specifications, sample trims, inference methods).
- ▶ Record the design, implementation status, and outcomes for each check in the reporting sheet.

## Referee report deliverable

- ▶ Use the referee-report tab inside the reporting sheet to summarize findings, robustness evidence, and recommendations.
- ▶ Focus on clarity: describe reproducibility outcomes, major/minor issues, and follow-up suggestions for the original authors.
- ▶ Keep evidence-linked: cite code cells, logs, or file names so the organizing team can audit quickly.

## Post-event follow-up

- ▶ Focus groups by treatment capture qualitative experience across arms.
- ▶ De-identified outputs enter the replication archive once the preregistration lock lifts.
- ▶ Participants receive summary results before journal submission and can provide feedback.

# ChatGPT Plus toolkit

- ▶ **Advanced Data Analysis.** Run Python, upload files, and produce figures or tables in chat.
- ▶ **Browsing & Deep Research.** Reach current sources with citations and credibility checks.
- ▶ **Custom GPTs & Agent Mode.** Tailor assistants and supervise multistep execution inside your workflow.



## Q and A

Thank you.