US-RSE 2026 · RESEARCH IN THE OPEN

STAMPED in Practice

Improving research skills through reproducible experiments

John A. Lee
Austin Macdonald
Cody C. Baker
Yaroslav O. Halchenko

Center for Open Neuroscience
Dartmouth College

Poster guide

Our experiment: iteratively improve STAMPED research skills by validating the scientific results and research objects they produce.

Make the improvement process inspectable.

Link the skills, data, runs, assessments and revisions, so others can evaluate both the science and how the guidance evolved.

01 · STAMPED

The seven principles

Self-contained	Reach all essential components from one project root.
Tracked	Identify every component and record changes and computational provenance.
Actionable	Provide executable procedures and reproduction instructions.
Modular	Separate components for reuse and recombination.
Portable	Version environment specifications and avoid undocumented host dependencies.
Ephemeral	Run in fresh, disposable environments built from recorded specifications.
Distributable	Make components persistently retrievable in their recorded state, with explicit licenses.

Each principle is a spectrum, supporting incremental improvement. STAMPED paper.

THE EXPERIMENT

Test the skills through the work

Do successive revisions improve scientific adequacy, reproducibility and consistency?

Run	Fresh agents address the same scientific question.
Assess	Independent judges inspect the research object and attempt replay.
Revise	Use recurring findings to improve and compress the skills.
Repeat	Test the next pinned revision with fresh agents.

Three iterations of testing and revision. In the final iteration, evaluate without further tuning.

THE STUDY IS ITSELF A RESEARCH OBJECT

Follow the evidence

Scientific task + inputs
↓
Guidance + tested revisions
↓
Reports + code + results
↓
Assessments + lessons

Inspect both the scientific result and the decisions that shaped the experiment through the repository map. Apply the STAMPED checklist to your own research object.

02 · SCIENTIFIC RESULTS

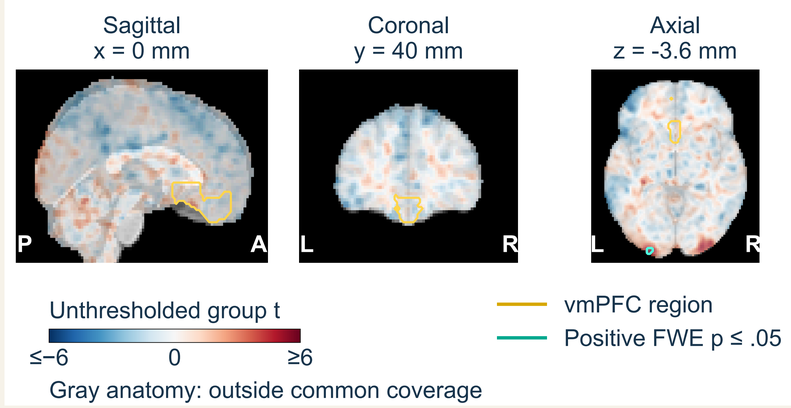
Replicating a gain-response analysis

Tom et al. asked how the brain represents potential gains and losses during risky choices.

Our replication: does potential gain increase ventromedial prefrontal cortex (vmPFC) activity? Agents test NARPS H1 using the same 54 equal-indifference AO86 participant maps.

NARPS paper · OpenNeuro data · AO86 maps · Scientific task

Group-level gain response

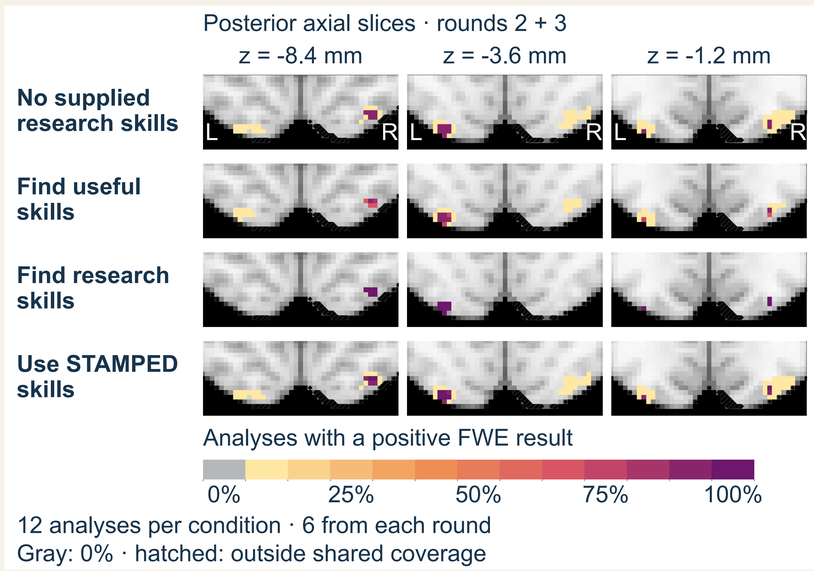


Round 2, run 09. One-sample test, 9,999 sign flips, whole-brain family-wise error (FWE) correction at 0.05. Analysis code · Numerical results

No corrected positive vmPFC signal

24 posterior voxels survived correction; none overlapped the analysis's Harvard–Oxford vmPFC proxy.

Where do repeated analyses agree?



Corrected findings vary across repeated analyses. Colour shows the fraction of 12 runs per condition reporting a positive FWE-corrected result in rounds 2 and 3. Each run retains its chosen method; 168,285 voxels share tested coverage. Figure code · Map inputs

03 · SKILLS-DEVELOPMENT RESULTS

Compare assistance across retained runs

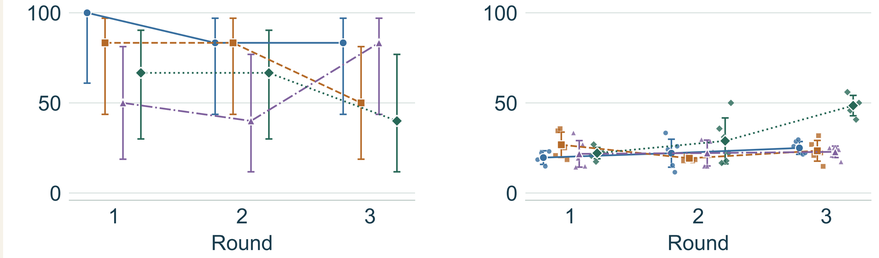
All conditions share the scientific task and software resources. The table names the added research-skill instructions; each linked mean opens that condition's checklists.

Rounds 1, 2, 3; 72 research runs. Pool all three rounds by condition. Comparison plan · Read a run

Instruction	Runs	Science	Replay	Checklist app mean met %
No supplied research skills	18	16/18	5/5	19.8% n=16
Find useful skills	18	13/18	2/3	19.5% n=18
Find research skills	18	10/17	3/7	19.0% n=17
Use STAMPED skills	18	10/17	3/6	27.7% n=17

Science: fully supported / assessed. **Replay:** all selected outputs matching / observed reruns. **App:** mean items met / all 31 (available n); the plot instead uses evaluated items. Scores + full responses.

How do outcomes change across rounds?



Science: 95% Wilson intervals, assuming independent runs. Checklist: individual scores and mean ± SD. Available n: science 5–6; checklist 4–6. Bars do not account for missing assessments or judge uncertainty. Figure code · Round data + denominators

NEXT STUDY

Future directions

Use stronger, calibrated judges and set clear completion rules for each round. Give every agent a common foundation in scientific tools and dependencies, and measure the coordinator effort supporting their work.

Test containerized isolation and more complete execution capture. Integrate skill-evolution tools and evaluate improvements on new scientific tasks.

Study design · Future study priorities

Use STAMPED. Try the skills. Bring us a study.

Share a hypothesis, accessible data and a reference result. Help us test the skills on your science—and discuss what better guidance should do.

Study + data

Use the skills

A reproducible study of reproducibility. Source, figure inputs and run-level evidence are linked from the study.

OpenAI Codex assisted analysis and poster preparation.

Explore CON

Dartmouth College
Psychological and Brain Sciences

Supported by ReproNim