The NIPS Experiment

Measuring Randomness in Peer Review Process

Prasoon Patidar

17-803 Empirical Methods | March 2026

Based on: Eric Price, "The NIPS Experiment" (Dec 2014)
https://blog.mrtz.org/2014/12/15/the-nips-experiment.html

Why study this?

THE PROBLEM

Researchers routinely experience what feels like arbitrary rejection, console themselves that the process is random, and then turn around and treat acceptances as genuine validation of quality.

THE GAP

Despite its outsized influence, the review process itself is almost never tested empirically. No one knows: if you reran the process, would you get the same results?

THE HOOK

If the process is substantially inconsistent, then career decisions built on top of it (hiring, tenure, funding) rest on unreliable foundations.

Experiment Design

Split

PC split into two independent halves

Duplicate

166 papers (10%) sent to both committees

Compare

Compare accept/reject for overlapping papers

Design choices:

- Binary outcome only — discards scores, confidence, review quality
- No control group for area chair effects or discussion dynamics
- "Fairness" rule (accepted if one side accepts) altered actual outcomes
    - https://berthuang.wordpress.com/2014/12/18/on-the-nips-experiment-and-review-process/

Same Data, Different Stories

43 / 166

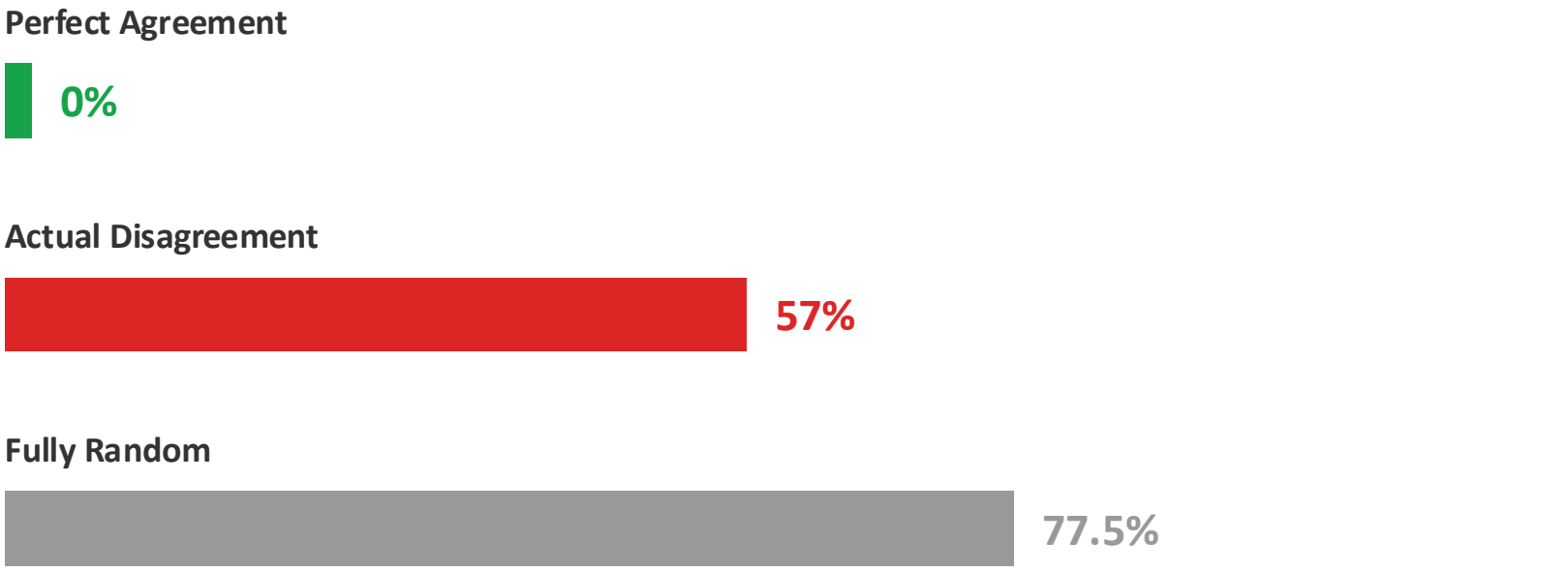
papers had different outcomes
across the two committees (25.9%)

~57%

of accepted papers were rejected
by the other committee

	Accepted by A Only	Accepted by B Only	Accepted by Both	Rejected by Both
Count	21	22	~16	~107
Disagreement Rate	**21 / 37 = 57%**	**22 / 38 = 58%**		

How Close to Random?



The actual reviewing process is much closer to random than to consistent. 95% CI: **40–75% of papers would flip on a re-run.**

Model 1: Messy Middle Model

- Not all papers are equally uncertain — splits them into three tiers:
  - **Clear accepts** — any committee would take them
  - **Clear rejects** — any committee would turn them down
  - **Messy middle** — outcome is essentially random

- Consistent with NIPS data if ~**50%** clear rejects, and rest are random

- Or: ~**7%** clear accepts (30% of accepted papers), and the other 93% random

The process reliably filters the extremes but is a lottery for everything in between.

Model 2: Noisy Scoring Model

- Each paper has a **true quality (v)** no one observes; 3 reviewers each produce a noisy score
- on the **average**: score > 6.5 accept, score < 6.0 reject, between = debate
- Two sources of variance:
  - $\sigma_{between}$ — real quality differences across papers (signal)
  - σ_{within} — reviewer disagreement on the same paper (noise)

When noise ≥ signal ($\sigma_{within} > \sigma_{between}$), which papers cross the cutoff is largely a coin flip.

What the Experiment Design Gets Right

- ✓ Real papers, real reviewers, real stakes — high ecological validity
- ✓ Clean identification: random assignment creates a natural counterfactual
- ✓ Unambiguous binary outcome avoids measurement ambiguity

Where the Experiment Design Falls Short

- ✗ Coarse measurement — binary outcome throws away rich score data
- ✗ Confounded noise sources — can’t isolate reviewer vs. area chair vs. discussion effects
- ✗ Non-independence — same reviewer community underlies both halves

What the Experiment Does Not Tell Us

1. Is the noise random or systematic?

Cannot distinguish unpredictable noise from structural bias against certain topics or approaches

2. Where does disagreement arise?

Reviewer scoring? Area chair aggregation? Discussion dynamics? The pipeline is a black box.

3. Does inconsistency mean the process is broken?

Expert disagreement on borderline work may reflect genuine ambiguity, not random error

Threats to Validity

- **Internal** — Committee halves share paper pool, norms, and reviewer networks. Area chair effects uncontrolled.
- **External** — Single conference, single year, single field. May not generalize to journals, grants, or other venues.
- **Construct** — Binary accept/reject is coarse. **Measures decision consistency, not evaluation consistency.**

Discussion

1. Is "consistency" the right metric for evaluating peer review?

2. How should this change how we evaluate researchers?

3. What would a more robust review system look like?

4. Should we run this experiment on graduate admissions? **:D**

Thank You

Price, E. (2014). "The NIPS Experiment." Moody Rd Blog.
https://blog.mrtz.org/2014/12/15/the-nips-experiment.html