

# From audit to action

Closing the transparency gap in quantitative social science

Matthew P. Robertson

Postdoctoral Fellow in Social Data Science, University of Mannheim

2025-10-20

# The paper

**Measuring transparency in the social sciences: political science and international relations**

*Royal Society Open Science*

Bermond Scoggins and Matthew P. Robertson

Published: 03 July 2024

<https://doi.org/10.1098/rsos.240313>

# Part 1: The study

# The problem: Science and trust

- The Royal Society's motto: *Nullius in verba* — “Take nobody's word for it”
- Yet much published social science research implicitly asks this of the reader
- Without data and code, readers cannot:
  - Verify computational accuracy
  - Check for errors
  - Test robustness to alternative specifications
  - Detect data falsification

# Two types of transparency problems

For observational studies with statistical inference:

- Need data and code to ensure computational reproducibility
- Verify results, check for errors, test robustness
- King (1995): “replication” in polisci usually means this

For experimental studies:

- Need preregistration to prevent selective reporting
- Distinguish confirmatory from exploratory analysis
- Combat HARKing, p-hacking, file drawer problem

# The replication crisis in psychology

What happened in psychology (2010s):

- Large-scale replication failures of published experiments
- Many Labs project: Much smaller effect sizes in replications
- Problem: Questionable research practices (QRPs) in experiments
  - Selective reporting of dependent variables
  - P-hacking and HARKing
  - File drawer problem

## Why it matters:

- False positives accumulate → inefficient interventions
- Field credibility eroded
- Good open science habits would have prevented many of these problems or enabled much faster detection and correction
- Transparency just keeps the field honest

# Terminology note

“Replication” can mean different things:

- **Psychology:** Often means collecting new data with same/similar design
- **Political science:** Often means computational reproducibility (reproducing results from same data/code)
- Political science conducts fewer experiments than psychology
- Nevertheless experiments we do conduct should be preregistered
- And observational studies need open data/code for computational reproducibility

# Background: the DA-RT statement (2014)

American Political Science Association initiative:

27 journal editors signed on to:

“increase the availability of data ‘at the time of publication through a trusted digital repository’”

“require authors to ‘delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to all relevant analytic materials’”

Implementation deadline: January 2016

# What do we know about transparency?

## Previous studies:

- Key (2016): 586 articles in 6 journals (2014-2015) → 58% have data+code
- Stockemer et al. (2018): 145 articles in 3 journals (2015) → 55% have data, 56% have code
- Grossman & Pedahzur (2020): 92 articles (Fall 2019) → claim “replicability utopia”

## Limitations:

- Small samples
- Time-intensive hand coding
- No comprehensive, field-level audit
- No systematic study of preregistration prevalence

# Research questions

Our two questions:

1. What proportion of papers using statistical inference make their data and code public?
2. What proportion of experimental studies were preregistered?

**Scope:**

- Top 160 political science and IR journals (Clarivate JCR)
- Years 2010-2021
- 109,000+ papers with accessible full text

# Methods

# The challenge

## Population of interest:

- Papers with statistical inference
- Papers with experiments

## Embedded in:

- All pol sci/IR publications
- Across 109k+ papers
- In 160 journals
- Over 12 years
- (crossref does not say *'this is a stats paper'* or *'this is a experimental paper'*)

# Phase 1: Data collection

1. Identify journals: Clarivate JCR top 100 (pol sci + IR) → 160 unique journals
2. Download metadata: Crossref API → 445,000+ papers
3. Filter by date: 2010-2021 → 93,931 papers
4. Obtain full text: PDFs and HTML → 109,000 accessible texts [*this part was hard*]
5. Convert to plaintext: `pdftotext` and `html2text` (plus OCR for PDFs where text extraction failed)

# Phase 1: Definitions

## **Statistical inference paper:**

Any paper involving mathematical modeling of data (OLS, regression, control variables, etc.)

Rationale: Mathematical modeling requires code. Without it, transformations cannot be exactly reproduced.

## **Experimental paper:**

Any article containing an experiment where researchers had control over treatment (original experiments only)

# Phase 1: Classification strategy

## Dictionary development:

- Read target papers
- Develop term dictionaries (data: 146 terms; stats: 437 terms; experiments: 89 terms)
- Iteratively refine
- Create document feature matrices

## Machine learning:

- Hand-coded training data (513 stats papers, 229 experiment papers)
- Trained Support Vector Machine and Naive Bayes classifiers
- Report SVM results (higher accuracy)
- SVM & NB performed better than `gpt4` circa '24

# Phase 2: Finding open data

Seven identification methods:

1. Query Harvard Dataverse API by journal → match to paper titles
2. Query Dataverse by paper title → match to paper titles
3. Extract Dataverse links from full text pdf/html → match to Dataverse
4. Download/scrape metadata to confirm data files exist
5. Extract “replication data” mentions + URLs from text → precarious data
6. Check other repositories (Figshare, Dryad, etc.)
7. For DA-RT journals: Validate by downloading article HTML and checking for data/code file extensions

# Phase 2: Finding preregistrations

Multiple search strategies:

1. Extract sentences mentioning “prereg”, “pre-analysis” + validate links
2. Search for references to OSF, EGAP, AsPredicted
3. Download all EGAP metadata, search corpus for registry IDs
4. Detective work: search author names in registries, match to papers

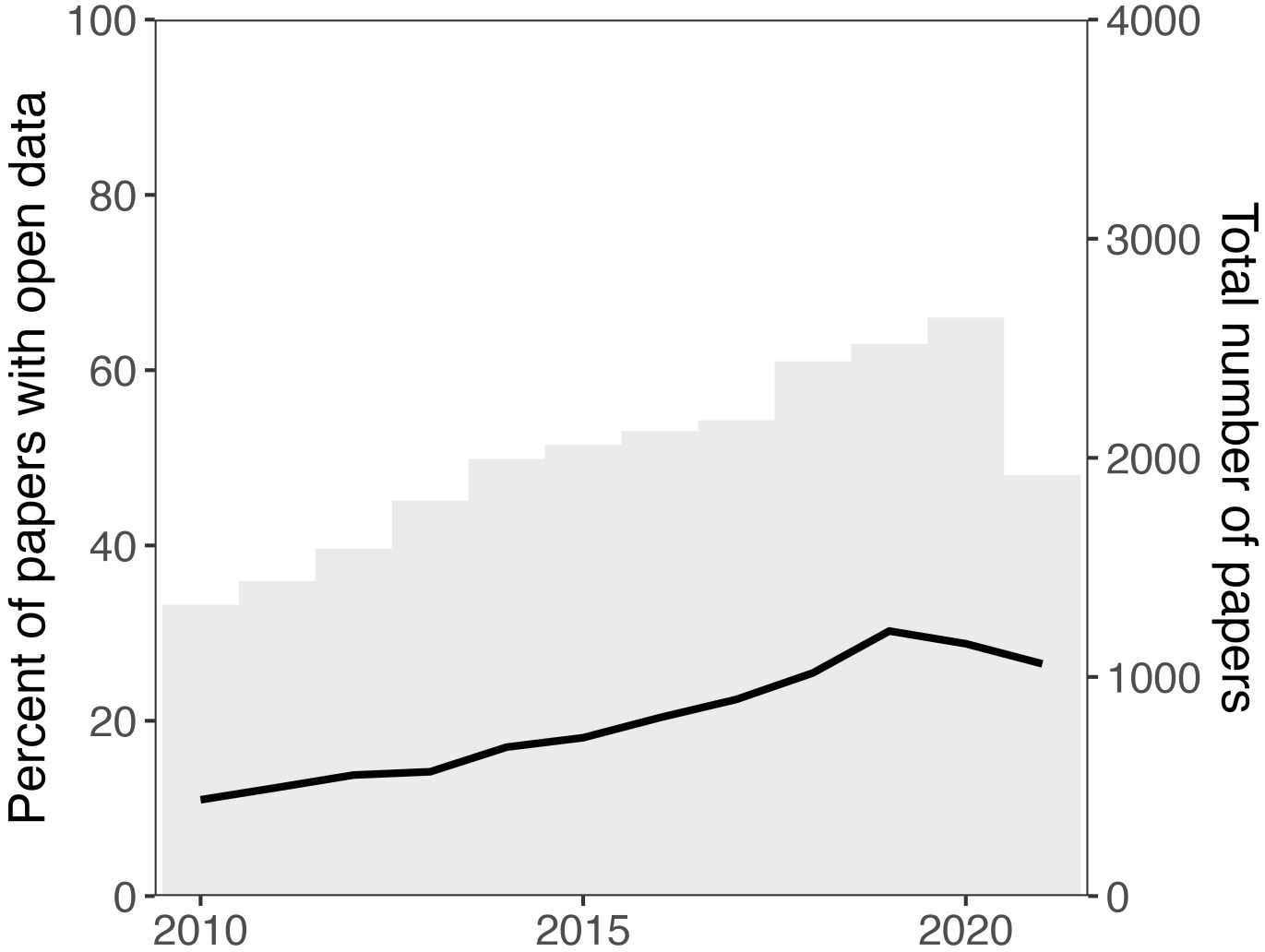
Note: We did not evaluate compliance with preregistration plans

This was time-consuming — many experiments don't mention their prereg in the paper!

Skipping many steps

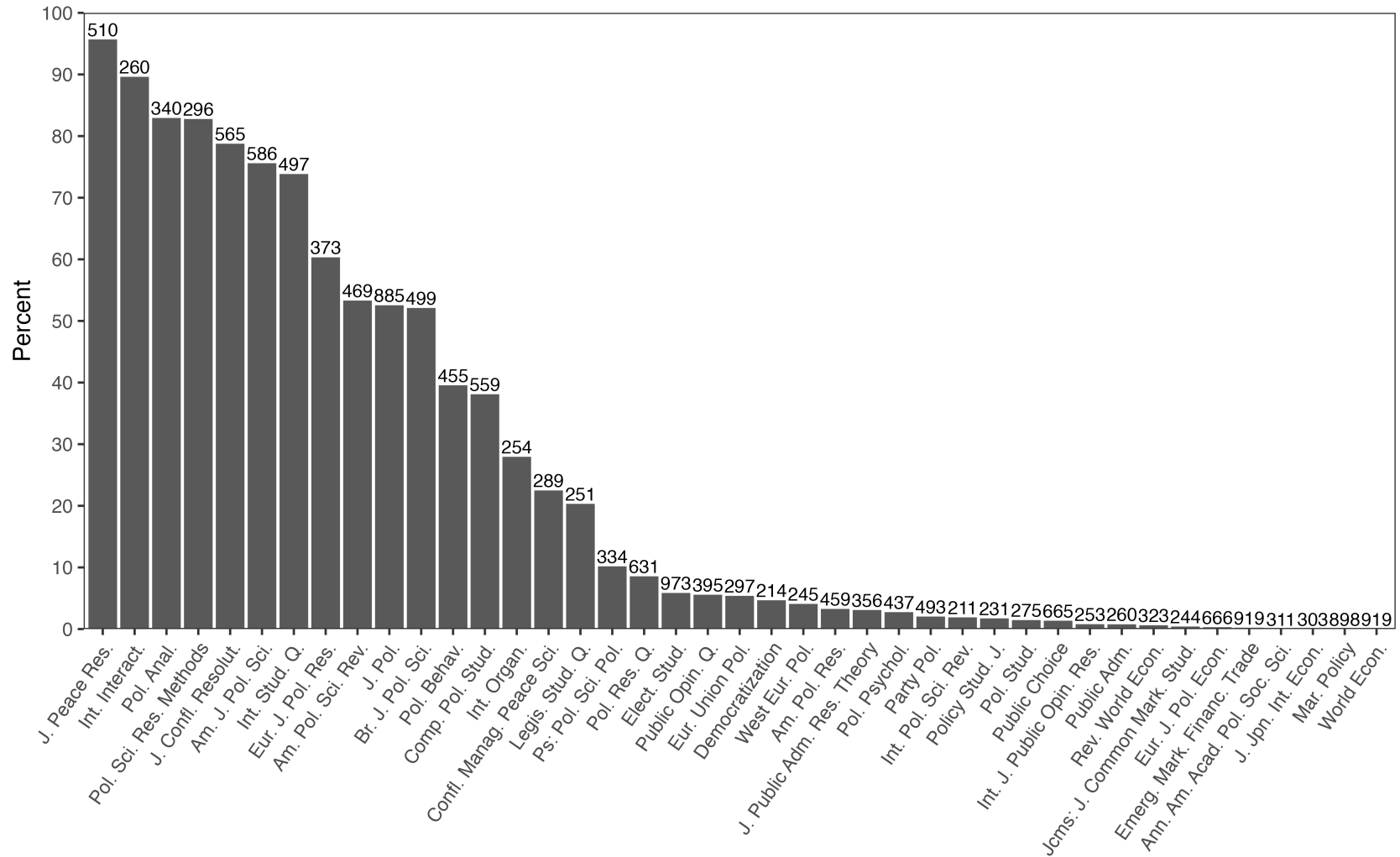
# Findings

# Open data over time



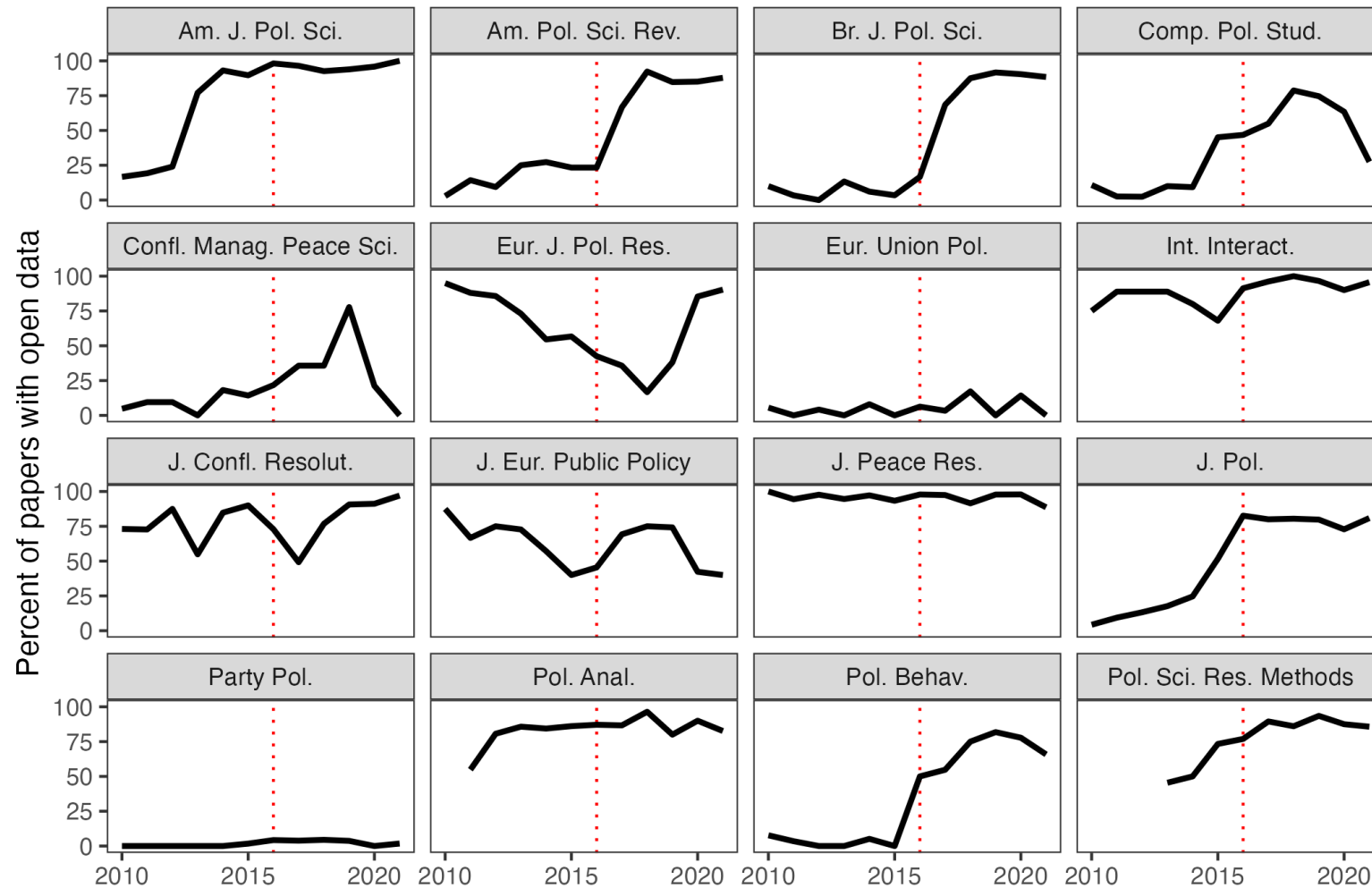
21% overall | Rising from 11% (2010) to 26% (2021)

# Open data by journal



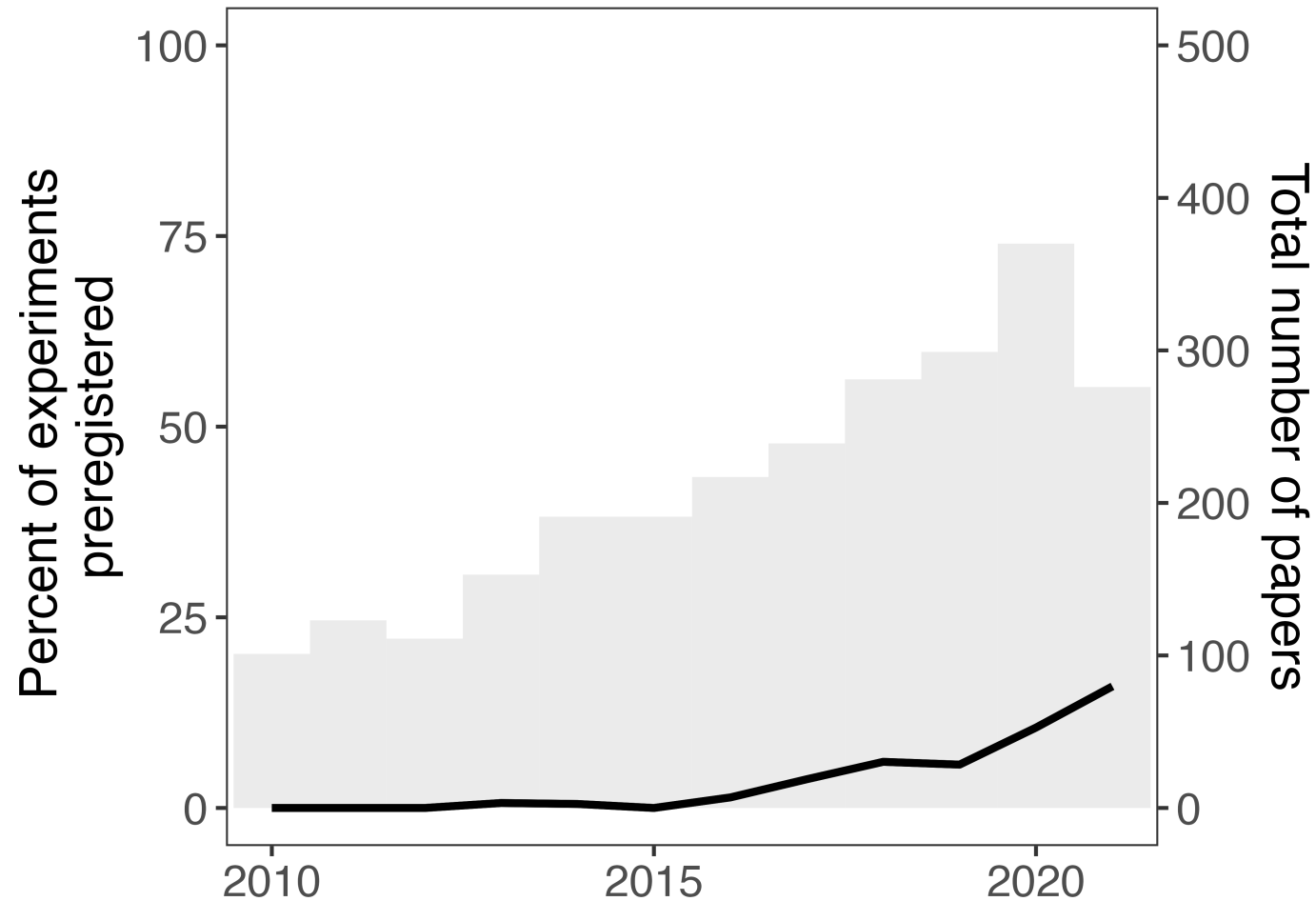
11 journals >50% open data | 25 journals <20%

# The DA-RT effect



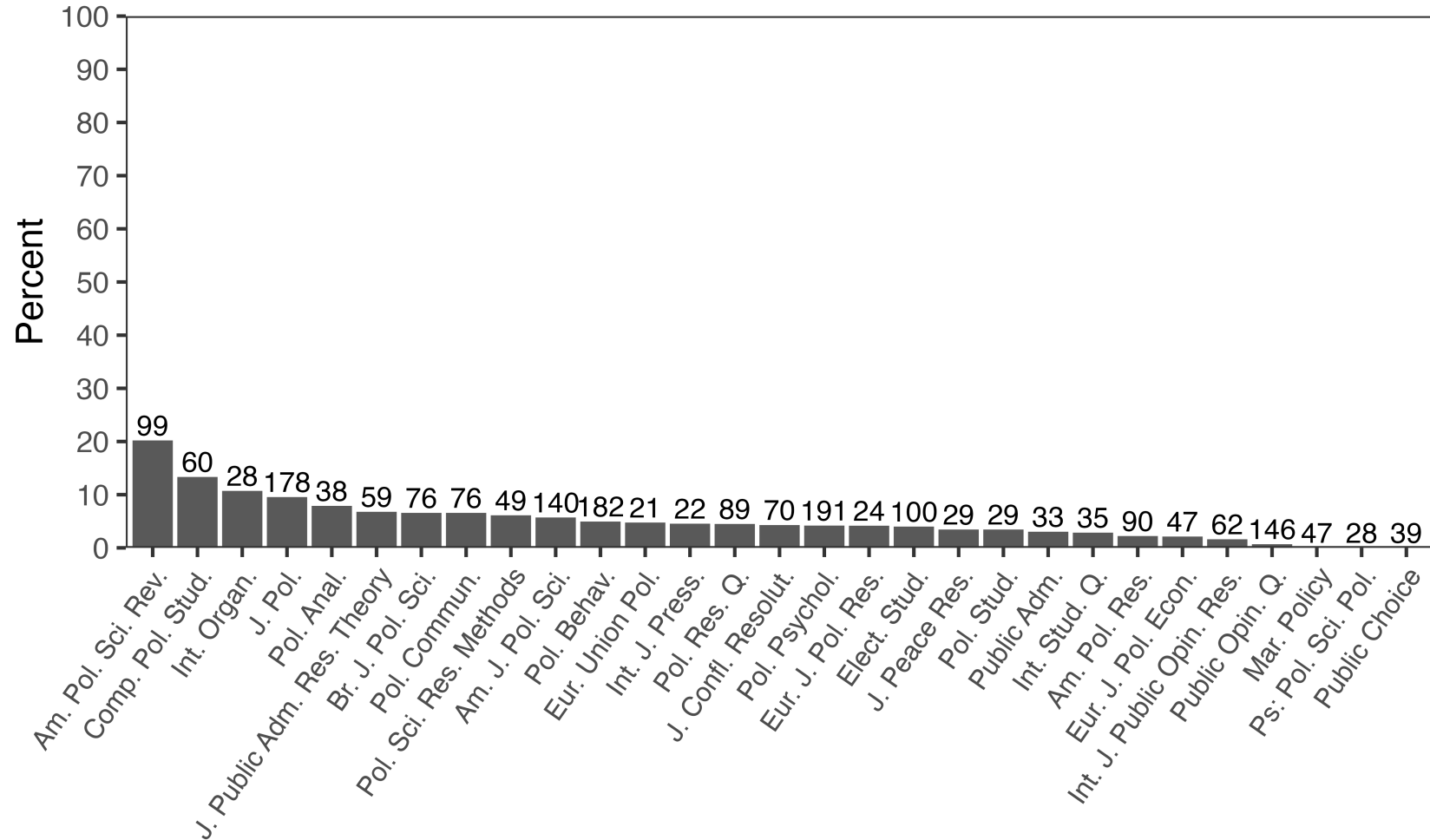
Journal policies work, but only when enforced

# Preregistration over time



5% overall | First in 2013 | Rising to 16% by 2021

# Preregistration by journal



Only APSR exceeds 20% | Most journals <10%

# Part 2: From audit to action

# The current gap

What some journals ask at submission:

- “Did you use quantitative data?”
- “Is this an experiment?”
- “Was it pre-registered?”
- “Data DOI?” (sometimes)

What reaches Crossref:

- Data DOI (sometimes, and then in wrong field)

**Problem:** No way to accurately assess field-level transparency without doing our study again and again

**Solution:** Just collect the data and send it to Crossref!

# What Crossref can hold

Crossref metadata has two relevant fields:

## 1. Assertions - Free-form structured claims (publishers define their own)

Nature (10.1038/s41586-024-07146-0):

```
1 {"name": "received", "value": "31 July 2023"},  
2 {"name": "Ethics", "value": "The authors declare no competing interests."}
```

eLife (10.7554/eLife.84364):

```
1 {"name": "peer_review_transparency", "value": "single anonymised"},  
2 {"name": "peer_review_published", "value": "review summaries, review reports..."}
```

## 2. Relations - Official Crossref recommendation for data/software linking

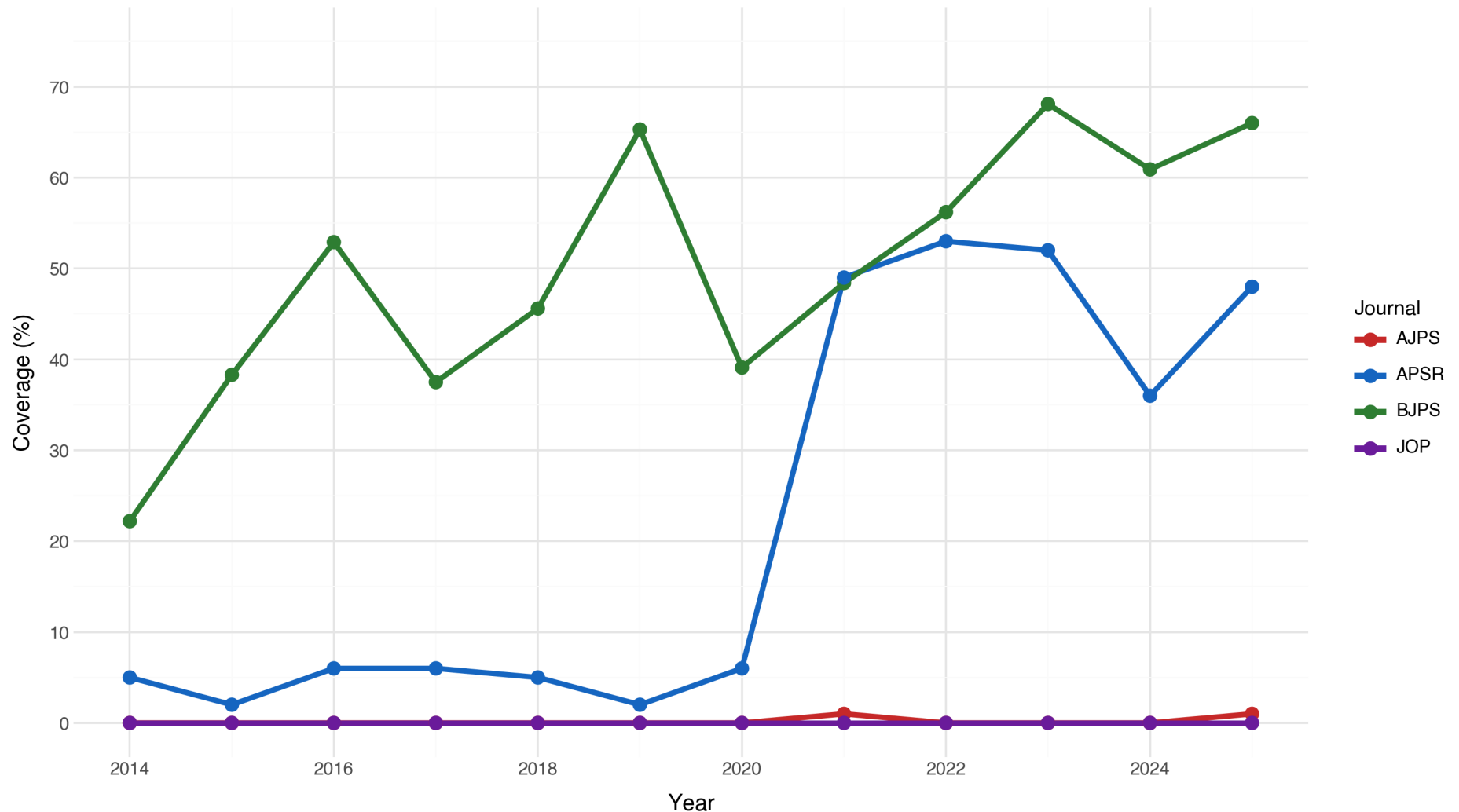
eLife (10.7554/eLife.84364):

```
1 "relation": {  
2   "has-preprint": [{"id": "10.1101/2022.10.17.512253"}],  
3   "is-supplemented-by": [{"id": "10.5281/zenodo.7944191"}]  
4 }
```

These can be crammed with anything the publisher wants

# BJPS and APSR are already doing half the job

Percentage of papers with Dataverse DOI in Crossref metadata (2014-2025)



BJPS: 11 years (50-68%) | APSR: 2021+ (~50%) | AJPS/JOP: 0%

# What's happening here?

- Journals use Editorial Manager or ScholarOne to handle submissions
- Authors answer “Did you use quantitative data?” and “Is this pre-registered?”
- Data DOIs and pre-reg links are collected later (*though for other journals not at all*)
- Editorial software generates Crossref XML and sends it to Crossref's servers
- In BJPS/APSR case, they are bundling the data doi in the references (wrong spot, but at least it's there)
- Most responses stay trapped in Editorial Manager's database
- But we can see *something* is working

# Remaining steps are easy

## Step 1: Update submission portal

- Add required fields: Research design, data DOI, pre-reg DOI, restriction reason
- Currently: Only some fields, not all required. Now we collect them all

## Step 2: Update publisher workflow

- Map portal fields → Crossref assertions & relations
- The journal would tell their software partner to update the crossref POST

## Step 3: #winning

- It will be in the crossref database
- Now transparency can be measured with a single database command (almost)

# A potential schema

```
1 {
2   "DOI": "10.1017/paper",
3   "assertion": [
4     {"name": "subfield", "value": "comparative_politics"},
5     {"name": "has_quantitative_data", "value": "yes"},
6     {"name": "has_experiment", "value": "yes"},
7     {"name": "preregistered", "value": "yes"},
8     {"name": "data_restriction_reason", "value": "confidential"}
9   ],
10  "relation": {
11    "isSupplementedBy": ["10.7910/DVN/XXXXX"],
12    "has-preregistration": ["10.17605/OSF.IO/XXXXX"]
13  }
14 }
```

- But questions arise: should epistemic status of experiment (exploratory / confirmatory) be mentioned? Absence or presence of prereg means different things in those cases.
- Much TBD

# General incentives

**Prestige:** Transparency leadership = journal reputation. Early adopters (BJPS, APSR) gain standing (?)

**Editorial efficiency:** Required portal fields = automatic enforcement. Less chasing authors.

**Field impact:** Political science as a field looks like a prestigious leader in prestigious practices

**Low cost:** One-time portal customization. Transparency dividends forever.

# Incentives for professional associations

Probably the impetus has to come from professional associations to journal editorial boards.

Associations (EPSA, APSA, MPSA, ISA etc.) should be incentivised by:

- Pursuing excellence for its own sake
- Gaining prestige for the field by being leaders in transparent practices

# Thank you!

## Contact:

Matthew P. Robertson [matthew.peter.robertson@uni-mannheim.de](mailto:matthew.peter.robertson@uni-mannheim.de)

Bermond Scoggins [bermond.scoggins@anu.edu.au](mailto:bermond.scoggins@anu.edu.au)