CSE 599K

Empirical Research Methods

Winter 2025

The Science in Computer Science

Today

- Paper discussion: Is Computer Science Science?
- Is Science objective?
- Evaluation frameworks
  - Ethics
  - Peer review
  - Artifact evaluation and Replication

Is Computer Science Science?

Is computer science science?

- CS = science, engineering, and mathematics.
- *“CS is a grab bag of tenuously related areas thrown together”*
- *“CS is not a science, and its ultimate significance has little to do with computers”*
- *“Computing is not a science because it studies man-made objects”*

- *“Most scientific fields have saturated”*
- *“Science will never again yield revelations as monumental as the theory of evolution, general relativity, quantum mechanics, …”*
- *“Has computer science already made all the big discoveries it’s going to? Is incremental progress all that remains?”*

- CS constantly forms new relationships with other fields => new fields.
- Overclaiming (empty promises) hurts the credibility of CS*.
- Is the scientific method applicable to CS?

* *Should computer scientists experiment more,* Tichy, IEEE Computer, 1998.

Should computer scientists experiment more?

1. Is computer science an experimental science?
2. What can we learn from the Knight-and-Leveson experiment?
3. Traditional scientific method isn't applicable.
4. The current level of experimentation is good enough (1998).
5. Experiments cost too much.
6. Demonstrations will suffice (proof of concept is good enough).
7. There is too much noise in the way (the easy way out).
8. Progress will slow.
9. Technology changes too fast.
10. You'll never get it published.
11. Feature comparison is good enough (comparison on paper or verbally).
12. Trust your intuition.
13. Trust the experts.
14. Flawed experiments (unrealistic assumptions etc.).
15. Competing theories (RISC vs. CISC, OO vs. functional programming).
16. Soft Science and Misuse.

Is Science objective?

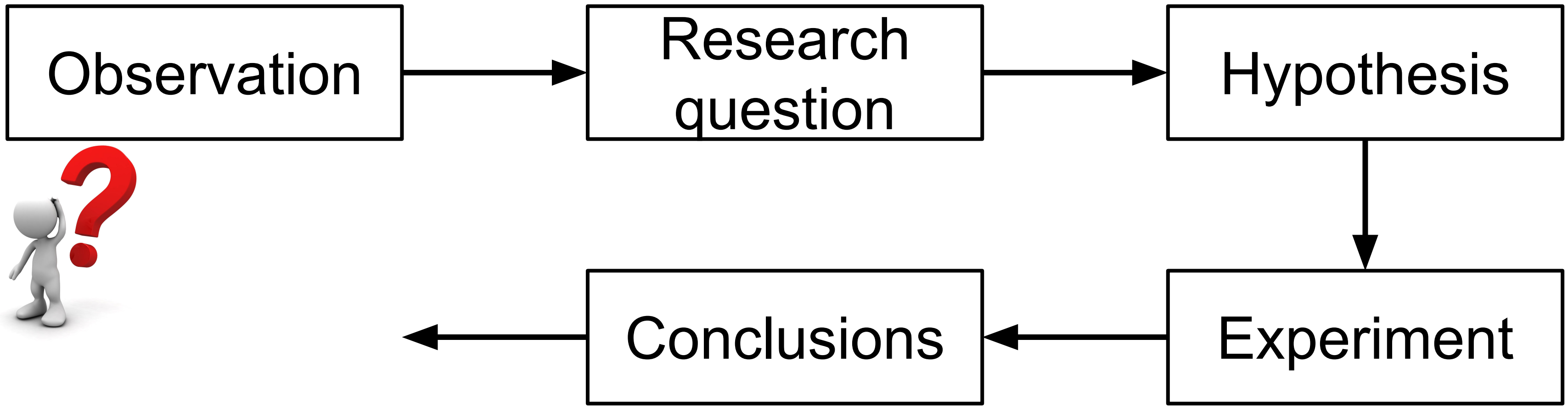
The holy grail: objectivity in science

Are falsifiability and NHST the solution?

The holy grail: objectivity in science

Are falsifiability and NHST the solution?

- Scientific method: rigorous framework and easy to execute



Observation

Research question

Hypothesis

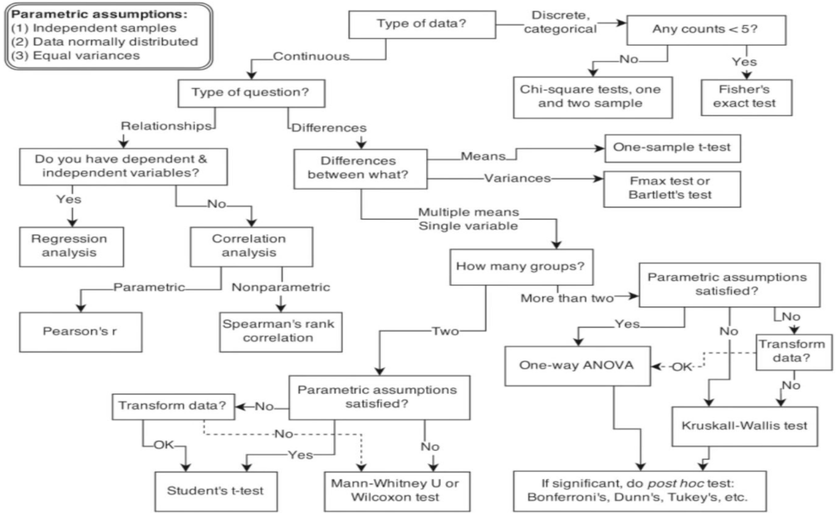
Experiment

Conclusions

The holy grail: objectivity in science

Are falsifiability and NHST the solution?

- Scientific method: rigorous framework and easy to execute
- Agreed-upon analysis methods and selection criteria

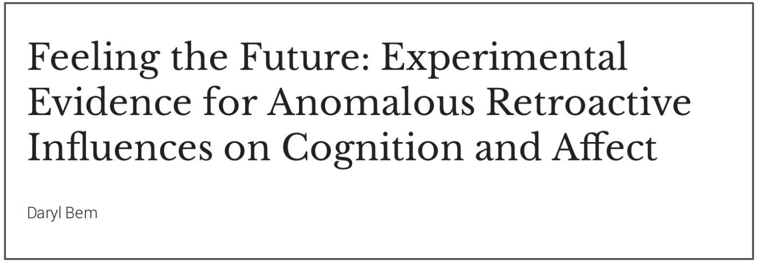


The holy grail: objectivity in science

Are falsifiability and NHST the solution?

- Scientific method: rigorous framework and easy to execute
- Agreed-upon analysis methods and selection criteria
- Mechanical and dichotomous decision making ($p<0.05$)

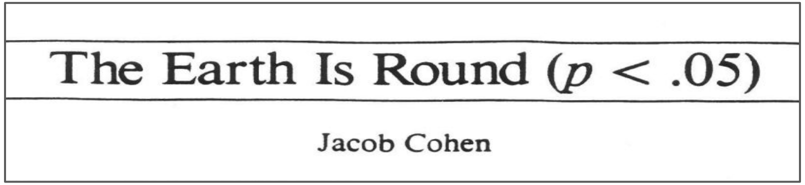
The holy grail: objectivity in science



Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

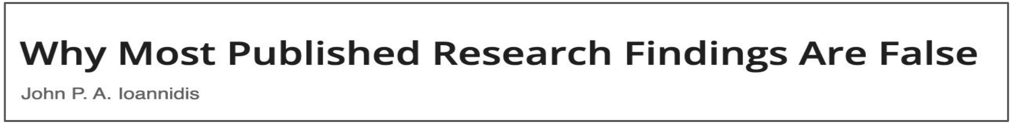
Daryl Bem

The holy grail: objectivity in science



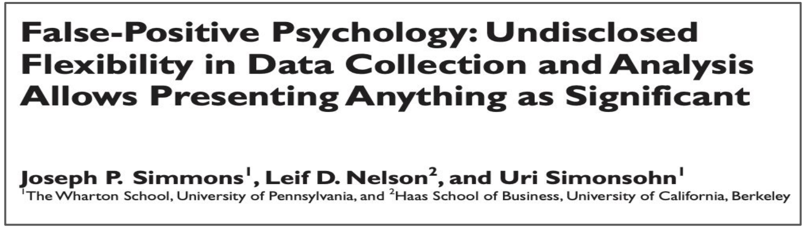
The Earth Is Round ($p < .05$)

Jacob Cohen



Why Most Published Research Findings Are False

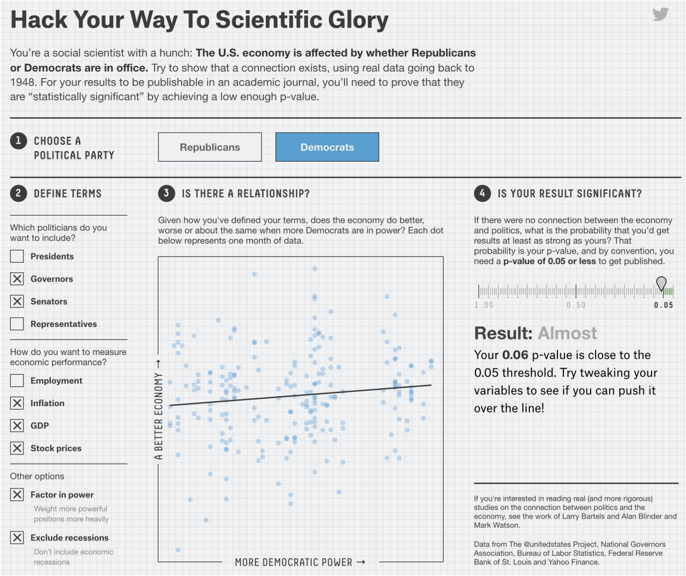
John P. A. Ioannidis



False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

The holy grail: objectivity in science



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

□ **Presidents**

☒ **Governors**

☒ **Senators**

□ **Representatives**

How do you want to measure economic performance?

□ **Employment**

☒ **Inflation**

☒ **GDP**

☒ **Stock prices**

Other options

☒ **Factor in power**
Weight more powerful positions more heavily

☒ **Exclude recessions**
Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

https://projects.fivethirtyeight.com/p-hacking

Has Science failed?

Ethical frameworks, transparency and replication go a long way

Science is subjective

Evaluation frameworks

Ethics

Core values (e.g., APA's ethics framework)

- Risks and benefits
    - Do benefits outweigh risks?
- Responsibility and integrity
    - Representation of a scientific field
    - Public trust
- Justice and fairness
    - No biased selection of control/treatment
- Rights and dignity
    - Awareness and consent
    - Privacy
    - Debriefing

Does not cover experiment design or data analysis.

Peer review

- Evolution and purpose (**grant funding vs. quality control** of published work).
- **Quality control vs. conclusion robustness** (peer review vs. replication).
- What are pros and cons for the **current peer-review process** (in your area)?

Peer review

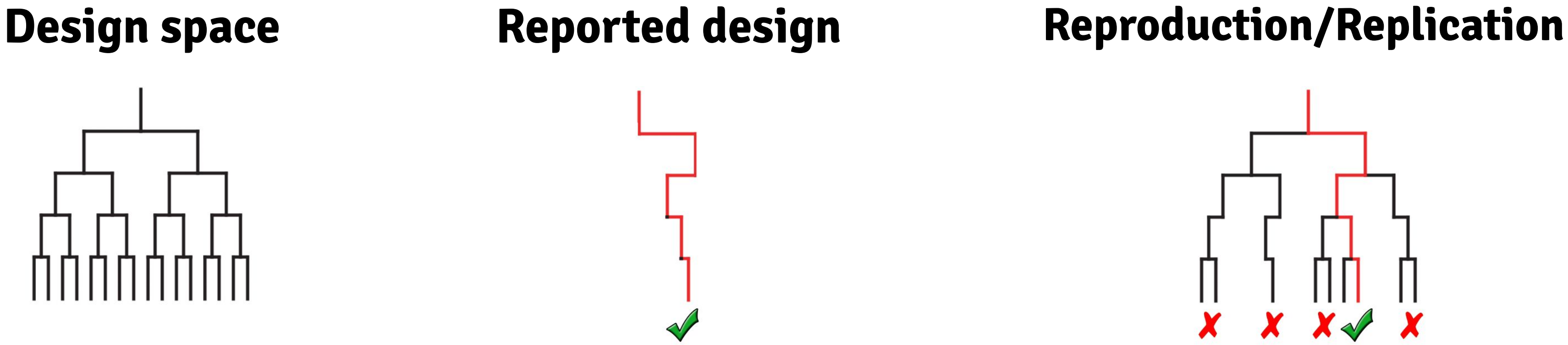
- Evolution and purpose (**grant funding vs. quality control** of published work).
- **Quality control vs. conclusion robustness** (peer review vs. replication).
- What are pros and cons for the **current peer-review process** (in your area)?

Latour defines ***science-in-the-making*** *as the* ***processes by which scientific facts are proposed, argued, and accepted****. A new proposition is argued and studied in publications, conferences, letters, email correspondence, discussions, debates, practice, and* ***repeated experiments****. It* ***becomes a "fact" only after it wins many allies*** *among scientists and others using it.* ***To win allies, a proposition must be independently verified*** *by multiple observations and there must be no counterexamples.*

Latour sees science-in-the-making as a messy, political, human process, fraught with emotion and occasional polemics.

Artifact evaluation and Replication

- Analysis grounded in a **conceptual model?**
- Clear **operationalization (implementation)?**
- **Implementation consistent with** the **model?**
- **Proper** use of **statistical methods?**
- Data interpreted in **context** of **prior knowledge?**
- Explored and validated **alternative hypotheses**?



Design space

Reported design

Reproduction/Replication

Artifact evaluation and Replication

- Analysis grounded in a **conceptual model?**
- Clear **operationalization (implementation)?**
- **Implementation consistent with** the **model?**
- **Proper** use of **statistical methods?**
- Data interpreted in **context** of **prior knowledge?**
- Explored and validated **alternative hypotheses**?

Transparency is key

- Transparent decision making (data collection and analysis)
- Shared instructions, data, and analyses (scripts)

Artifact evaluation and Replication

Artifacts Evaluated – Functional

Artifacts Evaluated – Reusable

Artifacts Available

Results Reproduced

Results Replicated

Transparency is key

- Transparent decision making (data collection and analysis)
- Shared instructions, data, and analyses (scripts)

https://www.acm.org/publications/policies/artifact-review-and-badging-current

Artifact evaluation and Replication

Artifacts Evaluated – Functional

Artifacts Evaluated – Reusable

Artifacts Available

Results Reproduced

Results Replicated

Transparency is key

- Transparent decision making (data collection and analysis)
- Shared instructions, data, and analyses (scripts)

What is the purpose of artifact evaluations?

https://www.acm.org/publications/policies/artifact-review-and-badging-current

Repeatability, reproducibility, and replicability

- **Repeatability**
  - Same research questions
  - Same experimental setup and artifacts
  - Same team
- **Reproducibility**
  - Same research questions
  - Same experimental setup and artifacts
  - Different team
- **Replicability**
  - Same research questions
  - Different experimental setup and artifacts
  - Different team

Note: the ACM defined replicability and reproducibility in the opposite way of most other scientific fields … now fixed!

Artifact badges

Artifacts Available
acm
Artifacts Evaluated
acm
Functional
Artifacts Evaluated
acm
Reusable

Pre-publication
(You)

Results Reproduced
acm
Results Replicated
acm

Post-publication
(Others)

Does the presence of a badge change your perception of a paper?

Artifact badges

	Repeated	**Reproduced**	**Replicated**
Team	*same*	*different*	*different*
Artifact	*same*	*same*	*different*

Artifact evaluations: the good, the bad, and the ugly

The good

- Lots of **sharing and transparency** (data availability is now an expectation).
- **Rose festival** and **reproducibility** (RENE) **tracks.**
- Some venues **invite replication studies** (as technical papers).

The bad

- **Artifacts** remain largely an **afterthought**.
- Lots of **overhead** (artifact eval) and **questionable focus** (reproducibility).
- **Little progress** on replicability.

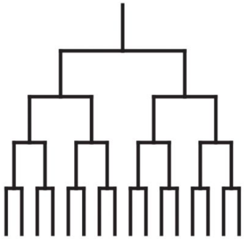
The ugly

- **Incentives**: Replicability isn’t valued.
- **False sense of security** (artifact vs. conclusions).
- **Specification crisis**: emphasis is on the implementation, not the design.

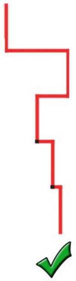
The role of peer review, artifacts, and replication

- Analysis grounded in a **conceptual model?**
- Clear **operationalization (implementation)?**
- **Implementation consistent with** the **model?**
- **Proper** use of **statistical methods?**
- Data interpreted in **context** of **prior knowledge?**
- Explored and validated **alternative hypotheses**?

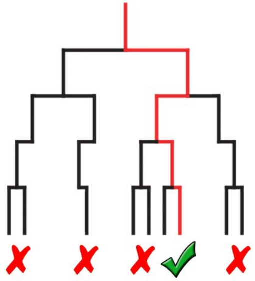
Design space



Reported design



Reproduction/Replication



Next week

- Quantitative vs. Qualitative research