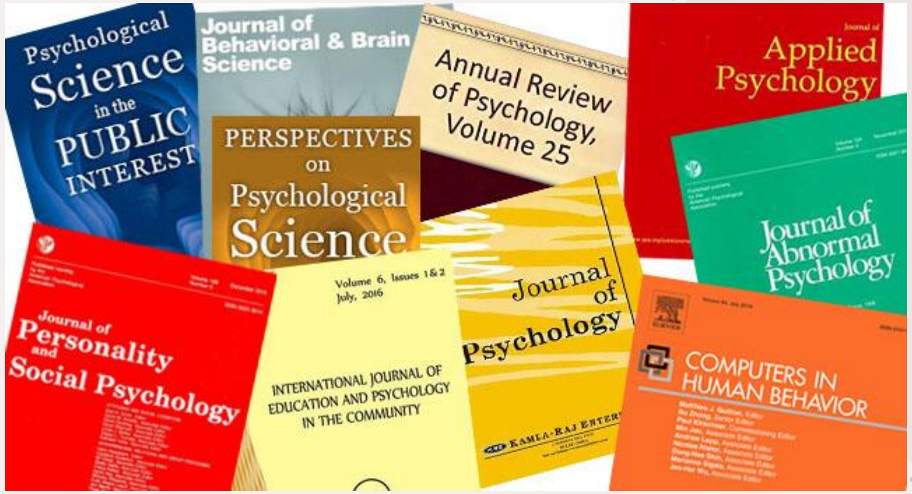
Open Science

PSY 612

The Goal of Science

- What do you think are characteristics of a good science?
  - **Objective:** minimize subjective bias of the researcher
  - **Generalizable:** results should generalize to more scenarios than the one specific experiment producing them
  - **Replicable:** the results of a study can be replicated by others



The 2010s in Psychology

- “The 2010s were considered psychology’s decade of crisis, revolution, or renaissance,” (Nosek et al., 2022)

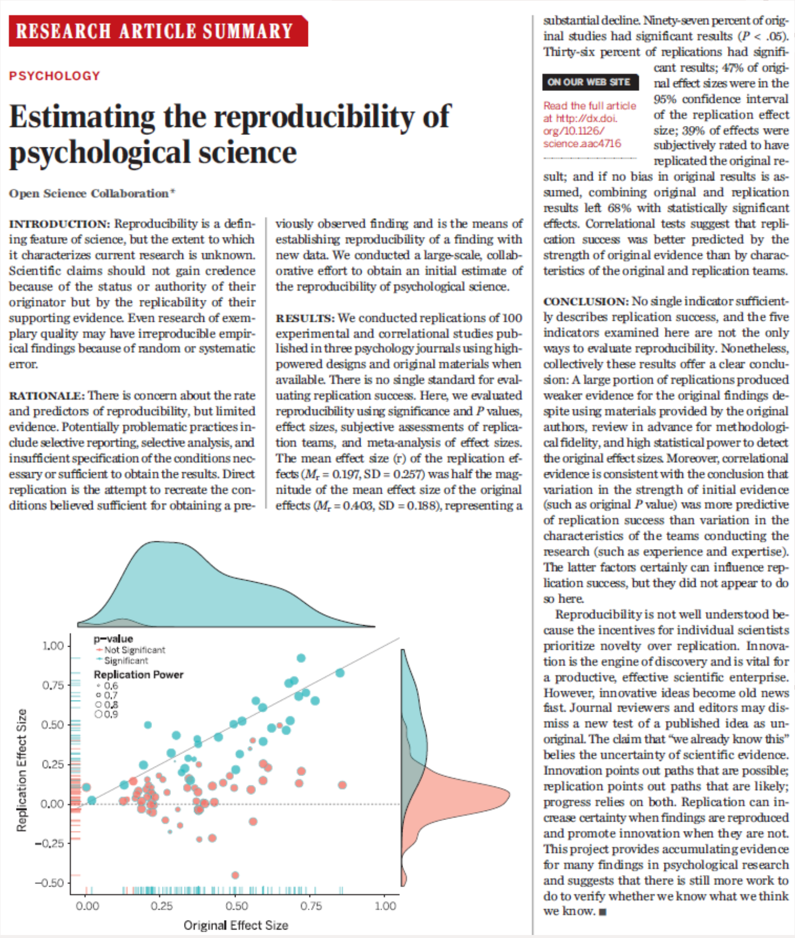
- For decades, common practices in psychology have been critiqued by methodologists:
  - Overemphasis on statistical significance ($p < .05$)
  - Publication bias
  - Inadequate statistical power
  - Lack of replication of published findings potentially due to questionable research practices

The Replication Crisis

- Open Science Collaboration (2015) attempted to replicate 100 published studies from three psychology journals using high-powered designs and the original materials (when available)

- <u>Results</u>: Only **36%** of the replications produced significant results, and the **average effect size across replications was half the magnitude** of the effect size found in the original studies

- Jeopardizes confidence in the psychological literature



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know. ■

Type I error

- When we perform research, we collect data from a **sample** of participants and attempt to estimate a model that generalizes to a **population** of people

- Data = Model + Error
  - Data: empirical observations on our predictor variable(s) and outcome from our sample data
  - Model: captures the **general relationship** between the predictor variable(s) and outcome
  - Error: the amount the model is off predicting people's actual scores
    - Systematic Error: There are systematic reasons for variations in people's scores on the outcome that could be explained with the addition of additional predictors
    - Random Error: the error that is **unique to the participants in your sample** that would not replicate in a new sample

When we overfit a model to the unique scores for the participants in our particular sample by making decisions informed by looking at the data, this increases the chances of Type I errors.

- Potentially not detecting a true effect, but rather, detecting a phenomenon that only occurs in the particular data set you are using (aka, misinterpreting *noise* as *signal*)

Questionable Research Practices (John et al., 2012)

- E-mailed 5,964 academic psychologists asking them to anonymously self-report whether they have engaged in the behaviors in table to the right
  - 2,155 people responses (36%)
  - The BTS (bayesian truth serum) condition was given an incentive to give honest responses

- Defensability ratings:
  - Was it defensible to have done so? (0 = *no*, 1 = *probably*, 2 = *yes*)

Table 1. Results of the Main Study: Mean Self-Admission Rates, Comparison of Self-Admission Rates Across Groups, and Mean Defensibility Ratings

Item	Self-admission rate (%)		Odds ratio	Two-tailed *p* (likelihood ratio test)	Defensibility rating (across groups)
	Control group	BTS group	(BTS/control)		
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5	1.14	.23	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0	1.08	.46	1.79 (0.44)
3. In a paper, failing to report all of a study's conditions	27.7	27.4	0.98	.90	1.77 (0.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5	1.57	.00	1.76 (0.48)
5. In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .054 is less than .05)	22.0	23.3	1.07	.58	1.68 (0.57)
6. In a paper, selectively reporting studies that "worked"	45.8	50.0	1.18	.13	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4	1.23	.06	1.61 (0.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0	1.45	.00	1.50 (0.60)
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	4.5	1.52	.16	1.32 (0.60)
10. Falsifying data	0.6	1.7	2.75	.07	0.16 (0.38)

False Positive Results (Simmons et al., 2011)

- Performed 15,000 simulations to examine how making flexible decisions after looking at the data affects the false-positive rate
  - Situation A: Analyzing different DVs and only reporting the results of the one that produces significant findings
  - Situation B: Collecting 20 observations per condition. If the result is non-significant, collect 10 additional observations per condition and re-test for significant
  - Situation C: Flexibly controlling for gender or for an interaction effect with gender to see if it makes the results significant
  - Situation D: Running multiple conditions and seeing how the inclusion, or exclusion, of any combination of conditions affects the significance of the results

- These are only a *small subset* of the types of flexible decisions that can be made by researchers after looking at data

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

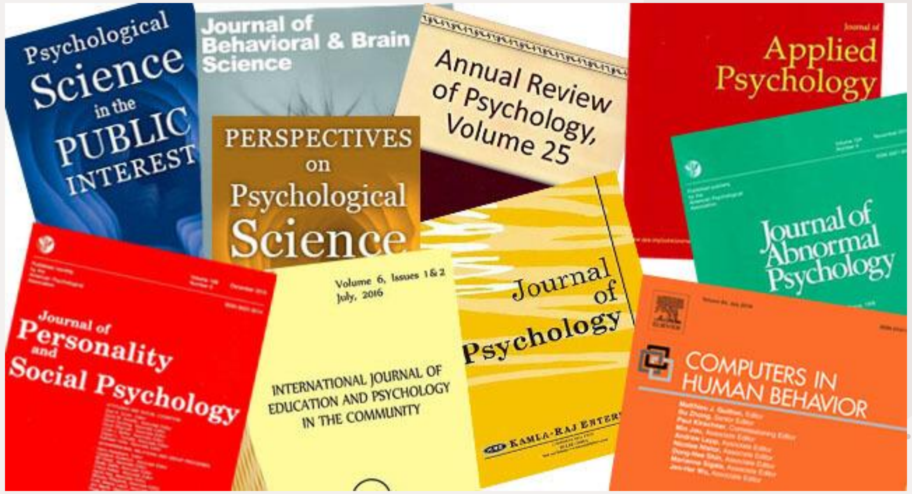
Making statistical decisions driven by observations of the data increases likelihood of Type I errors.

Publication Bias

- **Journals' publishing practices** have contributed to a culture that encourages *p*-hacking and QRPs
  - Only publishing novel and significant findings
    - Researchers discouraged from being honest about non-significant findings
    - Researchers discouraged from performing replication studies

- **Cultural practices** of establishing a **career** in psychology
  - "Publish or perish"
  - Career decisions, like hiring and attaining tenure, informed by whether researchers have published multiple studies in high-impact journals

- **Calls for reform** in publication practices and career practices



The 2020s in Psychology

- Open Science Initiatives
  - **Pre-registering** methods and analysis plans on Open Science Framework
    - Open Science Framework (OSF: https://osf.io/)
  - **Sharing data**, **code**, and **materials** to enable replication attempts
    - Over 100 journals have begun offering badges to authors engaging in open science practices (https://www.cos.io/initiatives/badges)
  - Increase in journals accepting **registered reports**
    - https://www.cos.io/initiatives/registered-reports

Confirmatory vs Exploratory Research

- **Confirmatory Research** (aka, Hypothesis-Confirming Research): collecting data to test a hypothesis decided on *a priori to* examining the data

- **Exploratory Research** (aka, Hypothesis-Generating Research): a *post-hoc* exploring of the data for unexpected patterns and discoveries
  - Can be followed up with a confirmatory study

- Both types of research are important, but researchers must **clearly distinguish** between them when reporting their results

Confirmatory Research

- Hypothesis testing
- Results are held to the highest standards
- Data-independent
- Minimizes false positives
- P-values retain diagnostic value
- Inferences may be drawn to wider population

Exploratory Research

- Hypothesis generating
- Results deserve to be replicated and confirmed
- Data-dependent
- Minimizes false negatives in order to find unexpected discoveries
- P-values lose diagnostic value
- Not useful for making inferences to any wider population

Pre-Registration

Publishing prior to examining data:

- Hypotheses
- Study Design
- Sample Size & Rationale
- Manipulated & Measured variables
- Planned Analysis, including specific models, decision criteria for excluding data, and plan for handling missing data

What is Preregistration?

When you preregister your research, you're simply specifying your research plan in advance of your study and submitting it to a registry.

Preregistration separates *hypothesis-generating* (exploratory) from *hypothesis-testing* (confirmatory) research. Both are important. But the same data cannot be used to generate *and* test a hypothesis, which can happen unintentionally and reduce the credibility of your results. Addressing this problem through planning improves the quality and transparency of your research. This helps you clearly report your study and helps others who may wish to build on it. For instructions on how to submit a preregistration on OSF, please visit our help guides.

PREREGISTERED

- Added benefit of thinking *carefully* and *critically* about your study before investing resources in it