How Can We Make Experimental Research Results More Reliable and Replicable?

John A. List, U. Chicago, ANU, and NBER
@Econ_4_Everyone

Building Confidence in (and Knowledge from) Experimental Results

1. Scientific research aims to create a stock of knowledge. Optimally adding to this stock requires confidence in the received estimates.
2. A key question concerning confidence revolves around the query: after a research finding has been claimed, what is the post-study probability that it is true?
3. Two unique features of the experimental approach situate it well to deepen the stock of scientific knowledge: selective data generation and the ability to enhance the notion, and role, of replications.

A Simple Bayesian Framework to Build Knowledge

$$PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

PSP: Probability that a declaration of a research finding, made upon reaching statistical significance, is true.

α: Level of statistical significance

1 - β: Level of power

π: Can think of this as the prior

Some Inference

Exhibit 15.5: PSPs With and Without a Statistically Significant Finding					
π	Power				
	0.20	0.30	0.50	0.70	0.80
PSP (reject null)					
0.01	0.04	0.06	0.09	0.12	0.14
0.05	0.17	0.24	0.34	0.42	0.46
0.10	0.31	0.40	0.53	0.61	0.64
0.20	0.50	0.60	0.71	0.78	0.80
0.30	0.63	0.72	0.81	0.86	0.87
0.40	0.73	0.80	0.87	0.90	0.91
0.50	0.80	0.86	0.91	0.93	0.94

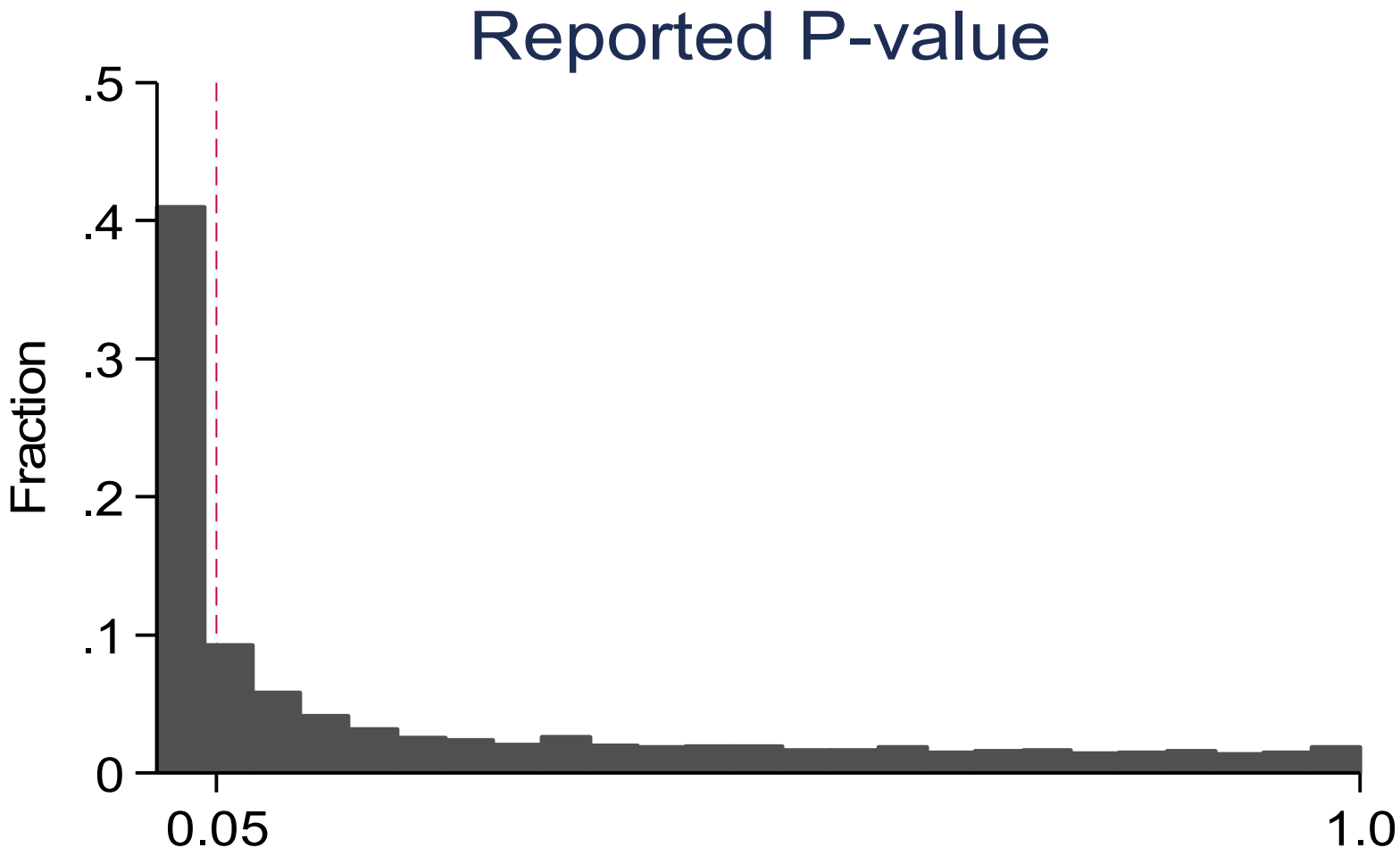
What Can Go Wrong?

Controlling the False Positive Rate

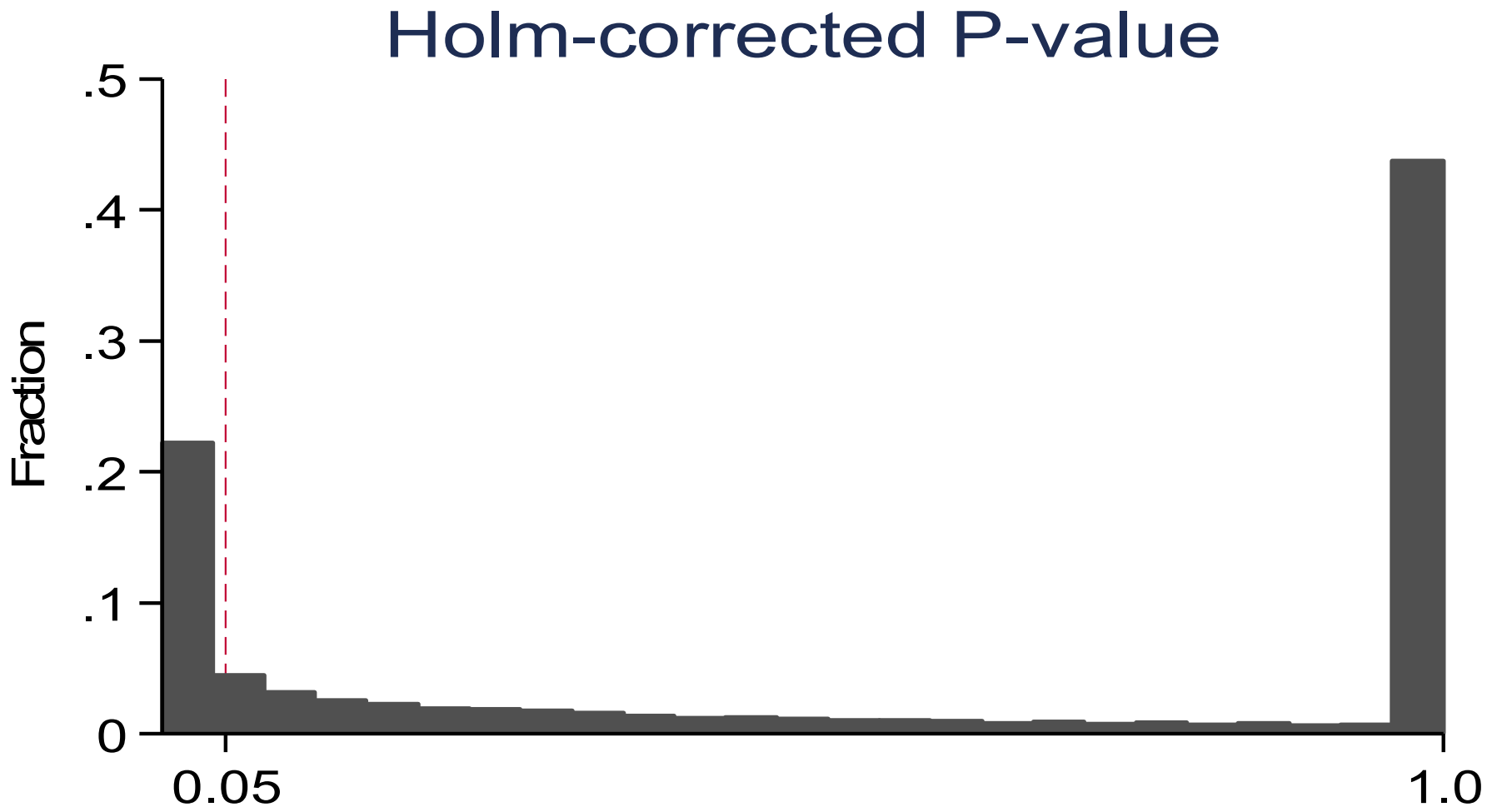
- Statistical Error (alpha)
- Human Error (how we generate/evaluate/interpret data)
- Human Fraud (less rare than we hope)

- Import of replication becomes clear

One Example of Human Error: MHT



Corrected P-values



Building Confidence: What Can We Do?

1) Reduce Bias

2) Promote Transparency

3) Promote Scrutiny

1. One Kind of Bias

Common belief is that significant results have much greater import than reporting null results.

- Scientific journals might prefer statistically significant "newsworthy" results
- Funders might reward scholars who produce noteworthy insights
- Ultimately, scientists might conclude that journal publications and streams of funding matter a great deal in tenure decisions

Yet, from a scientific perspective of building knowledge, such skewed preferences are flawed (see List, 2024).

Null Results Are Informative Too

- Let's return to our PSP equation and reverse engineer to find the inferential power of nulls:

$$PSP(Null) = \frac{(1-\alpha)(1-\pi)}{\beta\pi+(1-\alpha)(1-\pi)}$$

Exhibit 15.5: PSPs With and Without a Statistically Significant Finding					
π	Power				
	0.20	0.30	0.50	0.70	0.80
PSP (reject null)					
0.01	0.04	0.06	0.09	0.12	0.14
0.05	0.17	0.24	0.34	0.42	0.46
0.10	0.31	0.40	0.53	0.61	0.64
0.20	0.50	0.60	0.71	0.78	0.80
0.30	0.63	0.72	0.81	0.86	0.87
0.40	0.73	0.80	0.87	0.90	0.91
0.50	0.80	0.86	0.91	0.93	0.94
PSP (null result)					
0.01	0.01	0.01	0.01	0.00	0.00
0.05	0.04	0.04	0.03	0.02	0.01
0.10	0.09	0.08	0.06	0.03	0.02
0.20	0.17	0.16	0.12	0.07	0.05
0.30	0.27	0.24	0.18	0.12	0.08
0.40	0.36	0.33	0.26	0.17	0.12
0.50	0.46	0.42	0.34	0.24	0.17

Implications

- If our goal is to build scientific knowledge, then recognizing and rewarding null results, especially those that move priors, is important

- Side benefit: will reduce the level of bias in our science.

2. Promote Transparency

- **Pre-Registration (must be well timed)**
- **Pre-Analysis Plans (must be well timed)**
- **Registered Reports (not for all journals)**

Scientific transparency alone does not verify the validity of the received results. Rather, it permits an exploration of the received claims.

In this manner, transparency and scrutiny are complements in enhancing knowledge building.

Implications

- When building scientific knowledge it is important to understand that there is a crucial distinction between the probability that a reported significant finding in the literature represents a real relationship and the probability that an individual experiment has uncovered a real relationship.

- Side benefits of enhanced transparency: reduces bias and provides a better depiction of what the literature is finding.

3. Promote Scrutiny (Replications)

Pure replication: examine same question using the underlying original data set.

Robustness analysis: use the exact same data as the original analysis but modify the data or the empirical methods to see if the results are robust

Same population replication: running a new experiment closely following the original protocol to test whether similar results can be generated using random draws from the same underlying population.

Similar population replication: conducting an experiment with UCLA undergraduates to replicate a previous lab experiment conducted with University of Maryland undergraduates.

Disparate population replication: examining the same question and model using a population dissimilar from the original experiment.

Finally, the sixth and broadest replication category entails testing the hypotheses of the original study using a new research design; this as a **conceptual replication**,

How Fast Can We Build Confidence?

- Adapting PSP equation, we can explore power of one, two, and three same population replications:

$$PSP = \frac{\pi \sum_{i=r}^{n} \binom{n}{i} (1-\beta)^i \beta^{(n-i)}}{\pi \sum_{i=r}^{n} \binom{n}{i} (1-\beta)^i \beta^{(n-i)} + (1-\pi) \sum_{i=r}^{n} \binom{n}{i} \alpha^i (1-\alpha)^{(n-i)}}$$

The Power of Replication

Exhibit 15.6: PSP as a Function of the Power (β), Statistical Significance ($\alpha = 0.05$), Priors (π), and the Number of Successful Replications (i)								
	Power (1-β) = 0.80				Power (1-β) = 0.50			
	i = 0	i = 1	i = 2	i = 3	i = 0	i = 1	i = 2	i = 3
π	PSP				PSP			
0.01	0.14	0.72	0.98	1.00	0.09	0.50	0.91	0.99
0.02	0.25	0.84	0.99	1.00	0.17	0.67	0.95	1.00
0.05	0.46	0.93	1.00	1.00	0.34	0.84	0.98	1.00
0.10	0.64	0.97	1.00	1.00	0.53	0.92	0.99	1.00
0.20	0.80	0.98	1.00	1.00	0.71	0.96	1.00	1.00
0.30	0.87	0.99	1.00	1.00	0.81	0.98	1.00	1.00
0.40	0.91	0.99	1.00	1.00	0.87	0.99	1.00	1.00
0.50	0.94	1.00	1.00	1.00	0.91	0.99	1.00	1.00

What About Other Types of Replication?

Define "domain distance," ɖ, as the disparity in time, space, population, and the decision environment between the study setting and the setting of ultimate interest. We can then augment PSP equation (15.1) with a "domain distance" term ɖ:

$$PSP^{ɖ} = \frac{(1-\beta)\pi + \beta\pi ɖ}{(1-\beta)\pi + \beta\pi ɖ + [\alpha + (1-\alpha)ɖ](1-\pi)}$$

Exhibit 15.8: PSP^{d} with Various Distance (d) Levels

π	Power				
	0.20	0.30	0.50	0.70	0.80
d = 0.00					
0.01	0.04	0.06	0.09	0.12	0.14
0.05	0.17	0.24	0.34	0.42	0.46
0.10	0.31	0.40	0.53	0.61	0.64
0.20	0.50	0.60	0.71	0.78	0.80
0.30	0.63	0.72	0.81	0.86	0.87
0.40	0.73	0.80	0.87	0.90	0.91
0.50	0.80	0.86	0.91	0.93	0.94
d = 0.10					
0.01	0.02	0.03	0.04	0.05	0.05
0.05	0.09	0.12	0.17	0.21	0.23
0.10	0.18	0.22	0.30	0.36	0.39
0.20	0.33	0.39	0.49	0.56	0.59
0.30	0.45	0.52	0.62	0.68	0.71
0.40	0.56	0.63	0.72	0.77	0.79
0.50	0.66	0.72	0.79	0.83	0.85
d = 0.25					
0.01	0.01	0.02	0.02	0.03	0.03
0.05	0.07	0.08	0.10	0.12	0.13
0.10	0.13	0.16	0.19	0.23	0.25
0.20	0.26	0.29	0.35	0.40	0.43
0.30	0.37	0.41	0.48	0.54	0.56
0.40	0.48	0.52	0.59	0.64	0.66
0.50	0.58	0.62	0.68	0.73	0.75
d = 0.50					
0.01	0.01	0.01	0.01	0.02	0.02
0.05	0.06	0.06	0.07	0.08	0.08
0.10	0.11	0.12	0.14	0.15	0.16
0.20	0.22	0.24	0.26	0.29	0.30
0.30	0.33	0.35	0.38	0.41	0.42
0.40	0.43	0.45	0.49	0.52	0.53
0.50	0.53	0.55	0.59	0.62	0.63

Implications

- When building knowledge it is important to have rapid scrutiny so the course of science can quickly correct.

- We always focus on false positives but one might argue that researchers' tolerance for false negatives has potentially irreversible effects on the development of scientific knowledge: since false negative results are less likely to be followed up than false positives, self-correction is less likely to occur in these cases.

- Side benefits of scrutiny: reduces bias and provides a better depiction of what the literature is finding.

What Could Go Wrong?

- The Great Endangered Species!
  - Maniadis et al. (2017) survey experimental papers published between 1975–2014 in the top 150 journals in economics and estimate that the fraction of replication studies among all experimental papers in their sample is 4.2%.

- Changing Incentives

- ✓ Replications typically bring little recognition (few journals interested) and even induce scorn.
- ✓ JESA, JPE:Micro are promising steps.
- ✓ Need to change authors' incentives to collaborate with replicators. Should positive replications of one's work be considered a "super cite"?

Promoting Reproducibility

- Butera and List (2017): original investigators of a study commit to only publishing their results as a working paper and offer coauthorship of a second paper to others who are willing to replicate.

- Dreber et al. (2015) suggest using prediction markets with experts as quick and low-cost ways to obtain information about reproducibility.

Cites

Abrams, Eliot, Jonathan Libgober, and John A. List. 2020. “Research Registries: Facts, Myths, and Possible Improvements.” NBER.

Alevy, Jonathan, John List, and Wiktor Adamowicz. 2010. “How Can Behavioral Economics Inform Non-Market Valuation? An Example from the Preference Reversal Literature.” *National Bureau of Economic Research, Inc, NBER Working Papers* 87 (January). https://doi.org/10.3386/w16036.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2017. “Redefine Statistical Significance.” *Nature Human Behaviour* 2 (1): 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Butera, Luigi, Philip J. Grossman, Daniel Houser, John A. List, and Marie-Claire Villeval. 2020. “A New Mechanism to Alleviate the Crises of Confidence in Science-With An Application to the Public Goods Game,” Working Paper Series, , February. https://doi.org/10.3386/w26801.

Butera, Luigi, and John A. List. 2017. “An Economic Approach to Alleviate the Crises of Confidence in Science: With an Application to the Public Goods Game,” Working Paper Series, , April. https://doi.org/10.3386/w23335.

Cites

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36. https://doi.org/10.1126/science.aaf0918.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences* 112 (50): 15343–47. https://doi.org/10.1073/pnas.1516179112.

Levitt, Steven D., and John A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53 (1): 1–18. https://doi.org/10.1016/j.euroecorev.2008.12.001.

List, John A. 2004. "Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace." *Econometrica* 72 (2): 615–25.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104 (1): 277–90. https://doi.org/10.1257/aer.104.1.277.

———. 2015. "How to Make Experimental Economics Research More Reproducible: Lessons from Other Disciplines and a New Proposal." *Research in Experimental Economics* 18 (January). https://doi.org/10.1108/S0193-230620150000018008.

———. 2017. "To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study." *The Economic Journal* 127 (605): F209–35. https://doi.org/10.1111/ecoj.12527.

Tufano, Fabio, and John A. List. 2021. "On the Importance of 'Null Effects' in Economics." *Unpublished Manuscript*.