Assessing Replication: Lessons for Education Science

Jacob M. Schauer
jms@u.northwestern.edu
Institute for Policy Research
Northwestern University

IES Grant #R305B140042

SREE 2019

Outline

- Analysis methods for replication studies are still not a settled matter.
    - Analyses of replication studies in the social sciences have proceeded with some ambiguity, which has led to the use of methods with poor properties.
- We ought to approach the study of replication as (partially) a statistical problem.

Why should we focus on analysis methods?

Results of replication research is high-stakes/high-profile

- Replication is foundational to the logic and rhetoric of science:
  - "Non-reproducible single occurrences are of no significance to science." (Popper, 1959)
  - "Science advances on a foundation of trusted discoveries." (McNutt, 2014)
- If we can't re-create the effects of interventions found in experiments, how do we know they are effective?
- Recent replication research is published in high-impact journals and cited frequently.

An example: Open Science Collaboration (OSC)

Open Science Collaboration (2015)

- Attempted *direct replications* of 100 social/behavioral psych experiments
- Most attempts involved consultation with the original authors
- Determined that *61 of their 100 attempts failed*
- Published in *Science*
- Cited over 2,700 times in academic articles

OSC in the press

SCIENCE

How the GOP Could Use Science's Reform Movement Against It

The principles of openness, transparency, and reproducibility might be weaponized to defund and deny research.

ED YONG APR 5, 2017

OPINION | COMMENTARY

How Bad Is the Government's Science?

Policy makers often cite research to justify their rules, but many of those studies wouldn't replicate.

By Peter Wood and David Randall
April 16, 2018 5:56 p.m. ET

Half the results publis
ow a profe
esearcher
any influe
oducibilit
ns withou

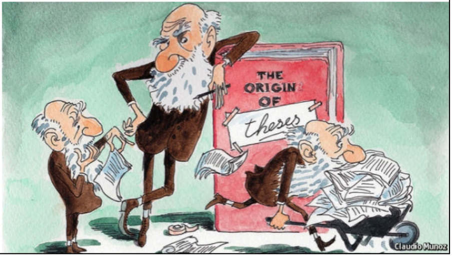
ADAM ROGERS SCIENCE 11.14.17 07:00 AM

THE DISMAL SCIENCE REMAINS DISMAL, SAY SCIENTISTS

Why bad science persists

Incentive malus

Poor scientific methods may be hereditary



What is the proper analysis?

Research programs note a lack of clear, standard methods:

- "There is no single standard for evaluating replication success," (OSC, 2015).
- "There are different ways of assessing replication, with no universally agreed-upon standard of excellence," (Camerer et al., 2016).

It has proven difficult to say what "replication" means:

- "Although we have an intuitive sense of what it means for results to replicate, the meaning becomes less clear the more closely we look," (Bollen et al., 2015).
- "The accomplishment of replication was dependent on contingent acts of judgment. One cannot write down a formula saying when replication was or was not achieved" (Shapin & Schaffer, 1985 re: Boyle and Huygens).

Formalizing analyses of replication as an applied statistics problem

Principles of applied statistics

1. What is it we're trying to measure?
   - What is a relevant operational definition of replication, and how can we translate that into a parameter?
2. What is the proper analysis method?
   - Most powerful tests
   - Estimates with low SE
3. What is the best way to collect data?
   - Sample size for required power, SE

Model (meta-analysis)

- $k \geq 2$ studies.
  - For the OSC, $k = 2$
- θ_i: effect of study i
  - θ_i may vary due to (possibly unknown) differences in experimental contexts.
- T_i: estimate of θ_i
- v_i: variance of the estimate T_i (e.g., due to sampling or randomization)
  - $v_i \propto 1/n_i$
- Assumption: $T_i \sim N(\theta_i, v_i)$
- Typically, one study T_1, v_1 is already conducted.

What is “replication”?

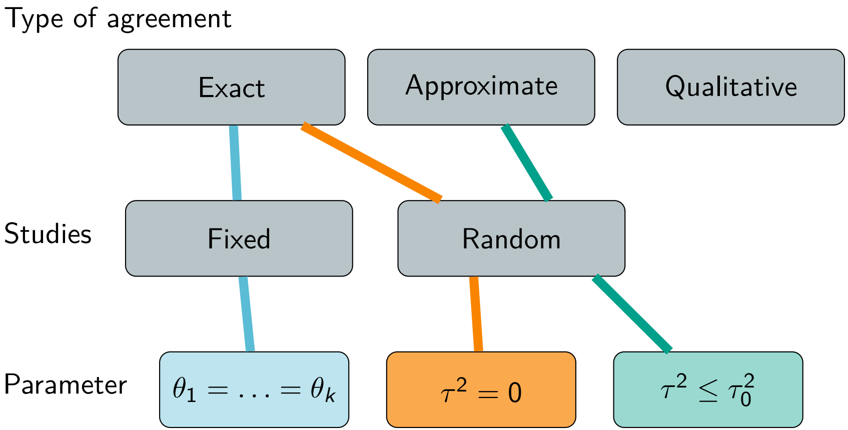
What does it mean for θ_i to be *the same*?

- Exact replication: $\theta_1 = \ldots = \theta_k$
- Approximate replication: θ_i are “practically the same”

Are we interested in *only* the k studies/effects?

- Fixed effects: the studies conducted are the only ones relevant to replication.
- Random effects: the studies conducted and their effects are sampled from some population.
  - θ_i are random draws from some distribution with:
    - $E[\theta] = \mu$, $V[\theta] = \tau^2$

Parametrizing "replication"



Example: confidence interval overlap

- Studies fail to replicate if T_1 is *not* in a 95% CI for θ_2.
- As a null hypothesis test:
  - H_0: $\theta_1 = \theta_2$
  - Test statistic $S = (T_1 - T_2)/\sqrt{v_2}$
  - Compare to a standard normal distribution
- Probability of saying studies failed to replicate when $\theta_1 = \theta_2$ is

$$1 - \Phi\left(\frac{1.96}{\sqrt{1 + v_1/v_2}}\right) + \Phi\left(\frac{-1.96}{\sqrt{1 + v_1/v_2}}\right)$$

- For OSC studies, this 25–40%!

Correction: Q test

Q test for exact replication is the UMP test.

1. Compute $Q = \sum_{i=1}^{k} \frac{(T_i - \bar{T}_\cdot)^2}{v_i}$
   - $k = 2 \implies Q = \frac{(T_1 - T_2)^2}{v_1 + v_2}$
2. Under H_0, Q has a chi-square distribution χ^2_{k-1}
   - $k = 2 \implies Q \sim \chi^2_1$
3. When H_0 is false, $Q \sim \chi^2_{k-1}(\lambda)$
   - $\lambda = \sum_{i=1}^{k} \frac{(\theta_i - \bar{\theta}_\cdot)^2}{v_i} \overset{k=2}{=} \frac{(\theta_1 - \theta_2)^2}{v_1 + v_2}$

Difference between effects

Increase power by decreasing v_2

Example

OSC replication of Payne et al. (2008)

- $T_1 = 0.753$, $v_1 = 0.0662$, $T_2 = 0.304$, $v_2 = 0.0229$
- $S = 2.96 > 1.96 \implies$ Failure to replicate
  - Probability of concluding replication failed when $\theta_1 = \theta_2$ is 32%
- $Q = 2.263 < 3.841 \implies$ Did not fail to replicate

Power of Q test

Was this test powerful? If not, what could they do differently?

- Power to detect failed replications depends on $|\theta_1 - \theta_2|$, and increases when v_2 decreases
- **Power to detect $|\theta_1 - \theta_2| = 0.5$ is 38%**
- **Even if $v_2 \to 0$, the power would only be 49%**
- **It is impossible to design a single replication of Payne et al. (and other OSC studies) to detect $|\theta_1 - \theta_2| = 0.5$ with much power.**
- This is because the power of the $k = 2$ design is limited by v_1.
- This limitation does not hold for $k > 2$ studies...

Discussion

- Applied statistics and meta-analysis provide a useful framework for approaching replication research.
- Choices about the proper definition and analysis will depend on the type of experiment, and the goal of the research.
- What are reasonable conceptions of replication we might want to study in education?

Thank you!

jms@u.northwestern.edu

Works Cited

- Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral,and Economic Sciences). National Science Foundation.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J.,Johannesson, M., ... Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436.

Works Cited (ctd.)

- McNutt, M. Reproducibility. *Science*, 343(6168), 229–229.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).
- Popper, K. (1959). *The Logic of Scientific Discovery*. Abingdon-on-Thames: Routledge.
- Shapin, S., & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton: Princeton University Press.