Studying replication: Lessons from applied statistics and empirical research

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

IES Grant #R305B140042
NSF Grant #1841075

JSM 2019
Session 486

Key points

- Analysis methods for replication studies are tricky.
  - 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.

Outline

1. What are replication studies (direct vs. conceptual)?
2. How can we define replication success/failure statistically?
3. What are the implications for existing methods and resluts?

What is a replication study?

Why replicate?

- Prove or falsify an existing finding.
- Examine sources of experimental variation:
  - Vary how an experiment is done
  - Conduct seemingly identical experiments

Direct vs. conceptual replications

Direct Replication

- Same procedure/protocol
- Same materials
- Same experimental units/population

Conceptual Replication

- Vary procedure/protocol
- Vary materials
- Sample from different population

Empirical research on replications

Major replication research programs in the social sciences attempted to:

1. Validate experimental protocol with original investigators, and/or
2. Standardize materials across multiple labs

Empirical research on replications

Study	# Exps	# Studies	Variation?	Falsification?
RPP/OSC	100	2		X
RPE	18	2		X
Many Labs	16	37	X	X
PPIR	11	11-17	X	X

Empirical research: Direct vs. conceptual replications

Direct Replication

- Same procedure/protocol
- Same materials
- Same experimental units/population

Falsify existing finding

Conceptual Replication

- Vary procedure/protocol
- Vary materials
- Sample from different population

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
roducibilit
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”?

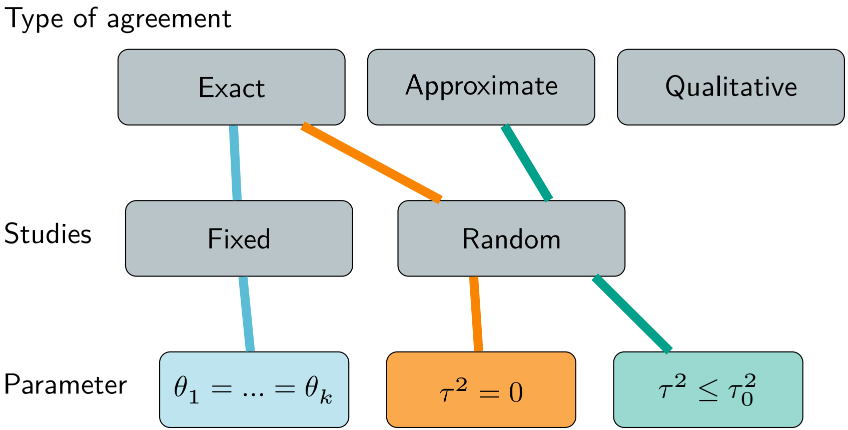
What does it mean for θ_i to be *the same*?

- Exact replication: $\theta_1 = ... = \theta_k$
- Approximate replication: θ_i are “practically the same”
- Qualitative replication: θ_i are the same sign: e.g., $\theta_i > 0$

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”



Properties of analyses of individual replications

Definition vs. analysis

- $F = \mathbf{1}\{\text{replication failure}\}$
  - $F = \mathbf{1}\{\theta_1 \neq \theta_2\}$
- $D = \mathbf{1}\{\text{analysis says replication failure}\}$
- It is possible that $D \neq F$

	$F = 0$	$F = 1$
$D = 0$	True success	False success
$D = 1$	False failure	True failure

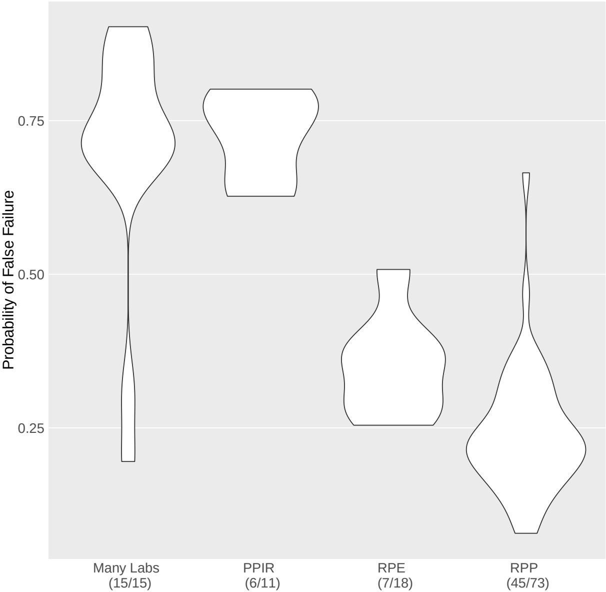
Example: confidence interval overlap

- 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 15–40%!

Probability of an error for “failed” replications



Correction: Q test

Q test for exact replication is the UMP test (a.k.a. prediction interval; Patil et al., 2017).

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, $Q \sim \chi^2_{k-1}(\lambda)$
   - $\lambda = \sum_{i=1}^{k} \frac{(\theta_i - \bar{\theta}_\cdot)^2}{v_i} \stackrel{k=2}{=} \dfrac{(\theta_1 - \theta_2)^2}{v_1 + v_2}$

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.**

Implications

1. Metrics to determine replication failure/success can be inaccurate (e.g., low power or uncontrolled type I error rate).
   - Schauer et al. (under review) show that averages of inaccurate individual determinations make for biased estimates of "failure rates."
2. It will often be impossible to design a replication (or several) in order to determine replication failure/success with high power.
   - The sensitivity of anlayses is limited by the design of the initial study (Hedges & Schauer, in press).
3. Potential re-framing of "replication" to mean:
   - Multiple studies obtain similar effects: consistency of $\theta_1, \ldots, \theta_k$ versus θ_1 is different from $\theta_2, \ldots, \theta_k$.
   - Power no longer limited by original study (Hedges & Schauer, 2019).

Thank you!

jms@u.northwestern.edu

Works Cited

Bollen, K., et al. (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.

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods.*

Hedges, L. V., & Schauer, J. M. (in press). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics.*

McNutt, M. Reproducibility. *Science*, 343(6168), 229–229.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science*, 11(4), 539-44.

Popper, K. (1959). *The Logic of Scientific Discovery*. Abingdon-on-Thames: Routledge.

Schauer, J. M., Fitzgerald, K., Peko-Spicer, S., Whalen, M., Zejnullahi, R., and Hedges, L. V. (Under review). The accuracy of large-scale analyses of replication.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton: Princeton University Press.

Replication research programs

- Camerer, C. F., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436.
- Klein, R. A. et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), 142-152.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).
- Schweinsberg, M., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66, 55–67.

Additional slides

Properties of aggregate patterns of replication

Proportion of failed replications

Replication crisis is about the prevalence of failed replications.

- N findings/experiments in a population.
- m experiments subject to replication attempts.
- $T_{ij}, v_{ij}, \theta_{ij}$ for study i of experiment j
- $F_j = 1$ if replication j failed.
  - $F_j = \mathbf{1}\{\theta_{1j} \neq \theta_{2j}\}$
  - $F_j = \mathbf{1}\{\theta_{1j}, \theta_{2j} \text{ different signs}\}$
- $D_j = 1$ if **analysis** determines replication failure.
  - $D_j = \mathbf{1}\{Q > c_\alpha\}$
  - $D_j = \mathbf{1}\{p_{1j} < 0.05, p_{2j} > 0.05\}$
- What we want is $\pi = \frac{1}{N}\sum_{j=1}^{N} F_j$
- Typically, what is reported is $\hat{\pi} = \frac{1}{m}\sum_{j=1}^{m} D_j$
  - 61% failure rate for the OSC.

Estimating π: sample selection

Are the m experiments representative of the population?

- It is often difficult to justify this (Gilbert et al., 2015).
- It is also unclear how to re-weight observations to minimize this issue.

If not, what about treating $m = N$ so that the sample is the population you care about?

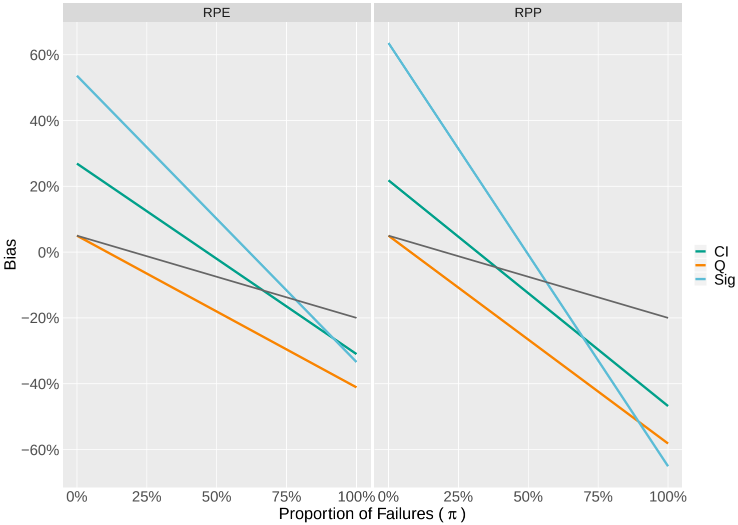
- $\pi = \frac{1}{m} \sum_{j=1}^{m} F_j$
- $\hat{\pi} = \frac{1}{m} \sum_{j=1}^{m} D_j$

Bias

$$E[\hat{\pi}] = \pi(\sum_{j:F_j=1} \underbrace{P[D_j = 1|F_j = 1]}_{\text{Prob. of detecting a true failure } (\beta_j)} + \sum_{j:F_j=0} \underbrace{P[D_j = 1|F_j = 0]}_{\text{Prob. of saying a successful replication failed } (\gamma_j)})$$

- Typically, β_j, γ_j vary depending on θ_{ij}, v_{ij}
- If $\beta_j = \beta$ and $\gamma_j = \gamma$, $\forall j = 1, ..., m$
  - Bias$(\hat{\pi}) = \pi(\beta - \gamma - 1) + \gamma$

Example: Bias



Alternatives

In multiple comparisons, methods for estimating π require procedures for D_j to have *controlled* error rates.

- Parametric mixture models (Tamhane & Shi, 2009)
  - of p-values: $p_j \sim \pi \underbrace{\text{Beta}(\alpha, \beta)}_{\text{failures}} + (1 - \alpha) \underbrace{U[0, 1]}_{\text{successes}}$
  - of test statistics: $\frac{T_{1j} - T_{2j}}{\sqrt{v_{1j} + v_{2j}}} \sim \pi \underbrace{N(\mu, 1)}_{\text{failures}} + (1 - \pi) \underbrace{N(0, 1)}_{\text{successes}}$
  - Require stronger assumptions and mixing components that are clearly separated.
- Moment estimator that is approximately unbiased (Storey, 2002).
  - RPP: $\hat{\pi} = 26\%$
  - RPE: $\hat{\pi} = 12\%$

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.
- Studying replicability via multiple replication studies to examine heterogeneity may allow for more sensitive analyses.
- Aggregating individual determinations about replication across programs that examine several experiments can be misleading.