Reproducibility in psych science

Estimating the reproducibility of psychological science, Science, 2015

Why replications?

- p-values<0.5 -> 5% false discoveries (FD), not bad?

Simple example why it could be much more than 5%:

- Assumption: 10% true effect, 90% null hypothesis
- type 1 error FD = .05*.9 = .045
- power = 80% -> discoveries = .8 *.1 = .08
- published discoveries = .08+.045 = .125
- false discoveries published = .045/.125 = .36

Lane and Dunlap 1978, Ioannidis, 2005; Pashler and Harris, 2012

Almost all the papers in psychological science claim "positive" findings

Sterling (1959), Bakker et al. (2012)

Why replications?

- p-values<0.5 -> 5% false discoveries (FD), not bad?
- Well, we already have a lot of replications:

Makel et al. (2012): Screening of 100 psych. journals since 1900

1. 1% replication efforts.
2. Of the 1%, only 18% *direct* replications
3. Of the 1%, only 47% done by 'new' investigators

- The authors:

Open Science Collaboration,

see Nosek, Persp Psyc Science, 2012

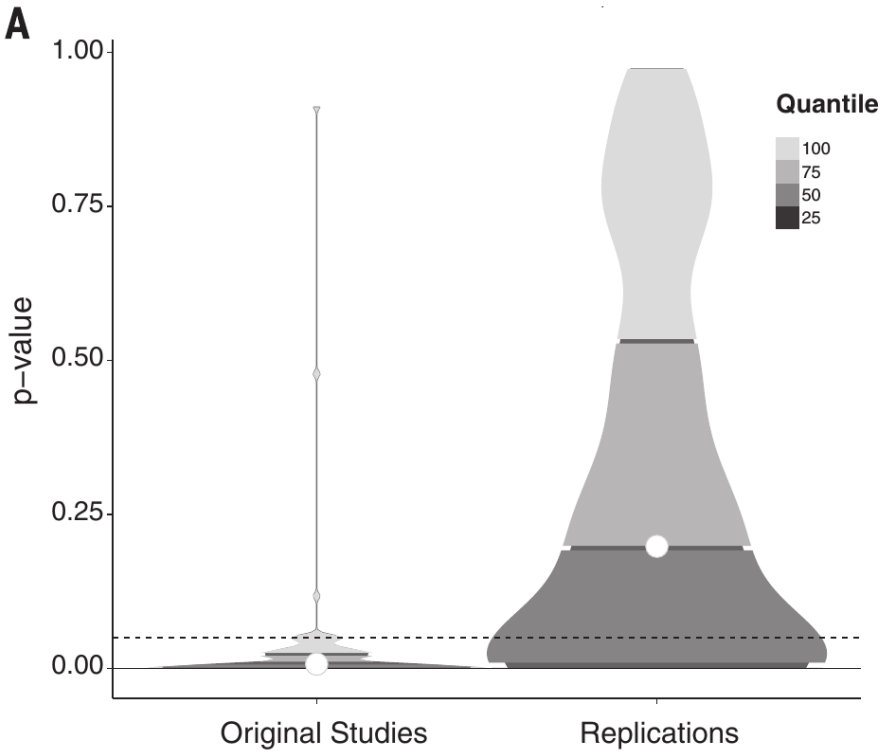
- Studies from 3 prestigious psych journals:
    - Psychological Science (PSCI), Journal of Personality and Social Psychology (JPSP), Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC)

- Not the first in Science, see AMGen study; Prinz, et al. 2011

Selecting the studies

- Provide a tractable sampling frame that would not plausibly bias reproducibility estimates
- Enable comparisons across journal types and subdisciplines
- Fit with the range of expertise available in the initial collaborative team
- Be recent enough to obtain original materials
- Be old enough to obtain meaningful indicators of citation impact
- Represent psychology subdisciplines that have a high frequency of studies that are feasible to conduct at relatively low cost.

p-values distribution

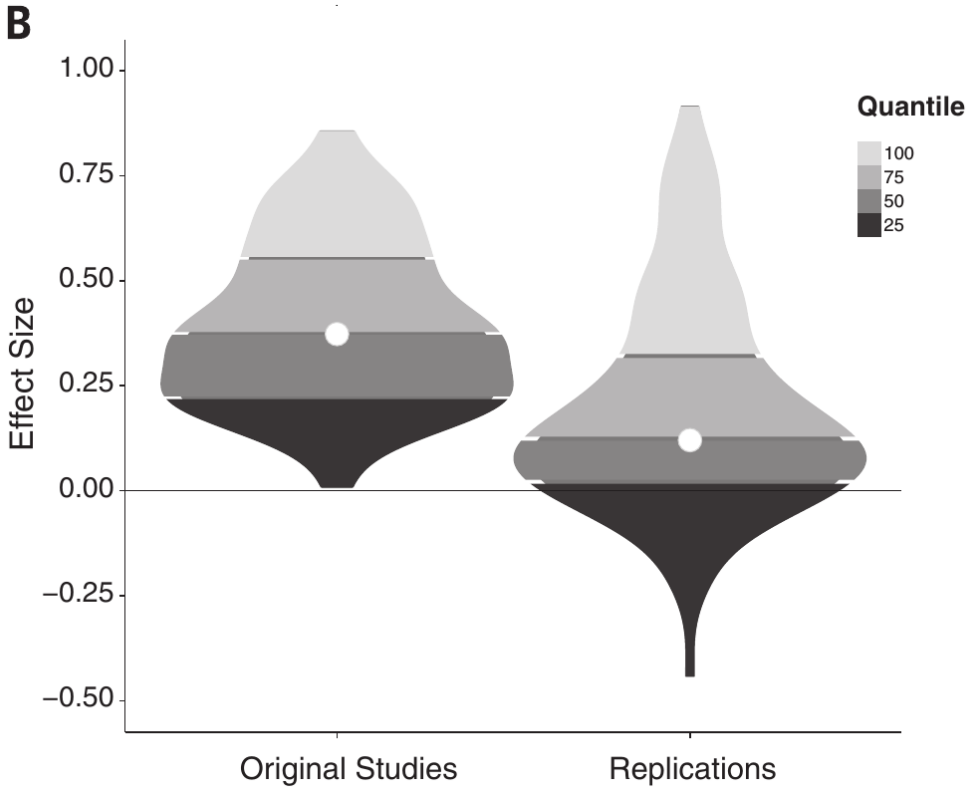


- **(Null distribution of p-values is uniform)**
- Replications' wide distribution against insufficient power

but

- significant and not-significant results are not (always) significantly different

Effect size distribution



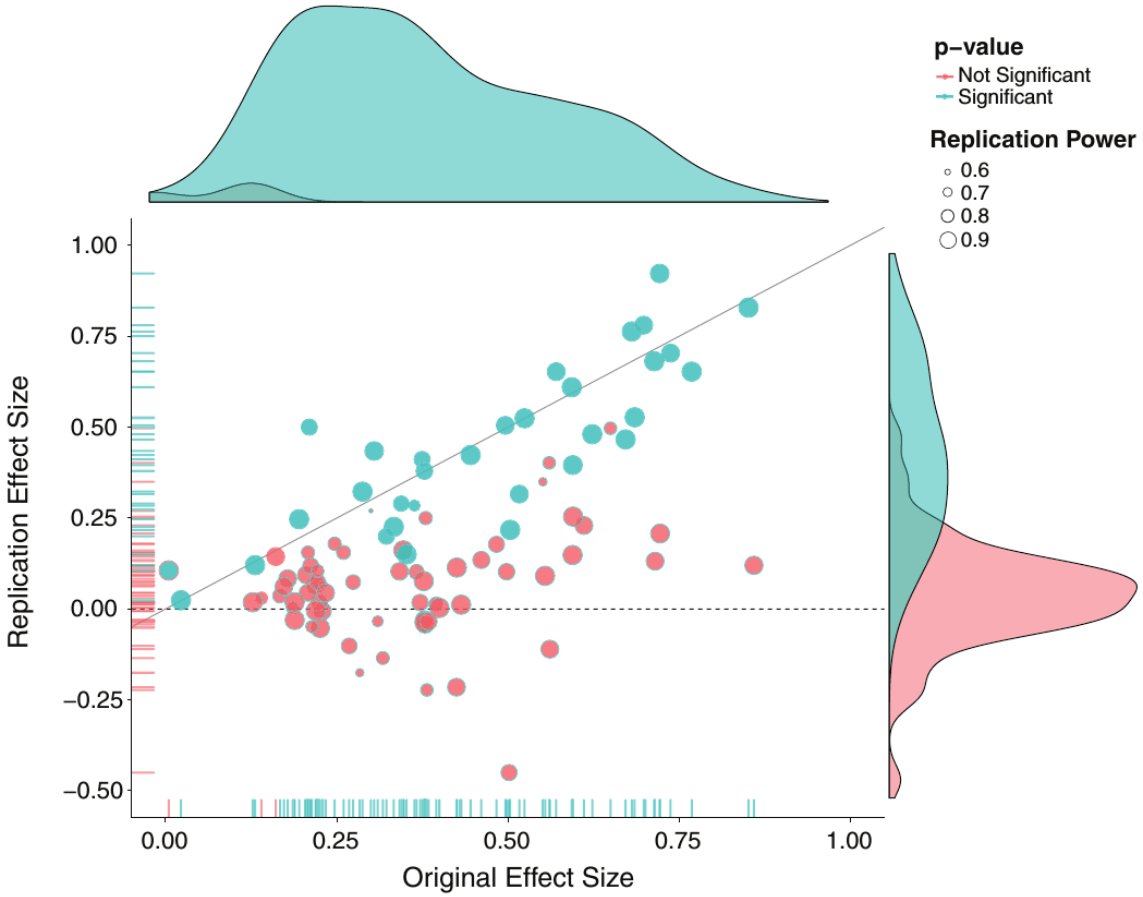
Post-Fisher: standard error a function of sample size

effect sizes -> correlations

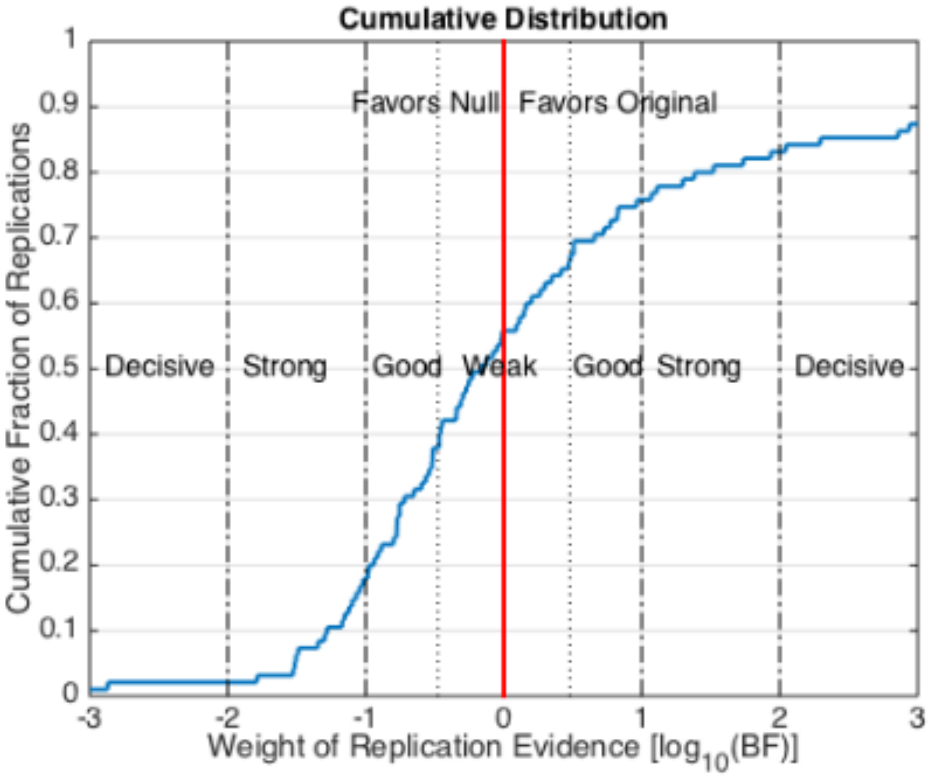
the original ES positive; the replication ES negative if opposite direction

"Replications with effects near zero but wide CIs get the same credit as replications that were bang on the original effect (or even larger) with narrow CIs" see Bayesian reproducibility project

Effect size *vs* p-values distribution



Bayesian factor



Not so dichotomous:

- Def.: success any evidence in favor of original effect (BF>1): 44% success
- Def.: success any moderate evidence in favor of original effect (BF>3): 34% success
- Def.: failure any moderate evidence in favor of the null (BF<1/3) 38% failure

The Bayes factor tells you which model (null or original published effect) the replication result is more consistent with, and larger Bayes factors indicate a better relative fit.

Bayesian reproducibility project

(almost) undebatable 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.

Possible causes

- Publication bias
- Experimenter bias
  - Selective reporting
  - Data selection/analysis
- Omitted variable bias
- Insufficient specification of the conditions necessary or sufficient to obtain the results
- Others statistical and sociological issues …

… there is room to improve reproducibility in psychology.

- Better powered studies
- Enhanced research standards including
  - Pre-registration of protocols (against 'file drawer')
  - Pre-registration of protocols, against p-hacking, p-fishing
  - Registration or networking of data collections within fields (as in fields where researchers are expected to generate hypotheses after collecting data)
  - Adopting from randomized controlled trials the principles of developing and adhering to a protocol.
  - Sample-size's increase
- Considering, before running an experiment, what they believe the chances are that they are testing a true or non-true relationship.
- Properly assessing the false positive report probability based on the statistical power of the test
- 21 Simmons words: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."
- Increase *direct*, instead of conceptual, replication
- Bayesian statistics (Wagenmakers, et al. 2011) instead of p-values/C.I.

Discussion time

Slides extra

Possible, but not probable, causes

- Experiences
- relevance

Intro

Measures and moderators:
features of the original study and replication as possible correlates of reproducibility:

1. publishing journal;
2. original effect size,
3. P value,
4. sample size;
5. experience and expertise of the original research team;
6. importance of the effect,
7. rated surprisingness of the effect.

We also assessed characteristics of the replication such as:

1. statistical power and
2. sample size,
3. experience and expertise of the replication team,
4. independently assessed challenge of conducting an effective replication, and
5. self-assessed quality of the replication effort.

Ioannidis' remedies:

1. Large studies where they can be expected to give very definitive results or test major, general concepts
2. Enhanced research standards including
   a. Pre-registration of protocols (as for randomized trials)
   b. Registration or networking of data collections within fields (as in fields where researchers are expected to generate hypotheses after collecting data)
   c. Adopting from randomized controlled trials the principles of developing and adhering to a protocol.
   d. Considering, before running an experiment, what they believe the chances are that they are testing a true or non-true relationship.
   e. Properly assessing the false positive report probability based on the statistical power of the test
3. Reconfirming (whenever ethically acceptable) established findings of "classic" studies, using large studies designed with minimal bias