Credit for the phrase “Psychology’s Renaissance” to Nelson, Simmons, and Simonsohn’s terrific 2018 Annual Review chapter.
Humility Statement

I’m a follower not a leader in methodological reform movement.

Not a statistician.

Come by my enthusiasm for reform by way of two decades of (mild) $p$ hacking of underpowered studies.

Many sub-issues open to debate and a diversity of views, no one-size-fits-all solutions.
CHALLENGES FACING PSYCHOLOGY

• Big
  – Develop into a cumulative science that is useful
    • Better theory, better measures, etc.

• Small
  – Stop publishing so many false positives and grossly exaggerated effect-size estimates
    • There are many rock-solid findings in psych
    • Bigger problem in some areas of psych than others
WHY BELIEVE PSYC OFTEN PUBLISHES FALSE POSITIVES AND EXAGGERATED EFFECT SIZE ESTIMATES?

• Most articles in scientific psych journals report NHST tests of relationship between variables in data and emphasize sig findings.
• Bem’s article reporting findings of extra-sensory perception
• Nosek et al. (2015) reproducibility (sic) project
• RRRs in PoPS and AMPPS
No reversals, and overall compelling evidence of a replicable effect, but it was sig in only a minority of the individ studies and overall size half that of the original.
<table>
<thead>
<tr>
<th>Project</th>
<th>Protocol Developed By</th>
<th>Final Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooler and Engstler-Schooler (1990)</td>
<td>Editors</td>
<td>Final Report</td>
</tr>
<tr>
<td>Hart &amp; Albarracín (2011)</td>
<td>Eerland, Sherrill, Magliano, &amp; Zwaan</td>
<td>Final Report</td>
</tr>
<tr>
<td>Sripada et al. (2014)</td>
<td>Martin Hagger</td>
<td>Final Report</td>
</tr>
<tr>
<td>Finkel et al. (2002)</td>
<td>Cheung, Campbell, &amp; LeBel</td>
<td>Final Report</td>
</tr>
<tr>
<td>Strack et al. (1988)</td>
<td>Wagenmakers, Beek, &amp; Dijkhoff</td>
<td>Final Report</td>
</tr>
<tr>
<td>Srull and Wyer (1979)</td>
<td>Bruno Verschuere and Ewout Meijer; Randy McCarthy and John Skowronski</td>
<td>Final Report</td>
</tr>
<tr>
<td>Many Labs 2</td>
<td>Klein et al.</td>
<td>Preprint Report and Commentaries</td>
</tr>
</tbody>
</table>
Many Labs 2: Klein et al. (2018)

28 studies replicated 60+ times each with decent N

15 obtained sig effect in predicted direction.

9 obtained ns difference in opposite direction.

Average ES estimates much smaller than original.
Other examples of influential reported effects recently challenged as difficult to replicate

- Power posing
- Social priming of a variety of sub-types
- Effects of the colour red
- Ego Depletion


Morey, Glenberg et al. (2018) many labs test of “action-sentence compatibility effect”

Ill effects of using notebook versus pen to take class notes.
Potential Reasons for Failures to Replicate

• Even ideal replications of non-trivial effects should fail $1-\beta$ of time. If non-trivial effect is smaller than initial finding then power may be lower than intended.

• Probably some replication attempt didn’t re-create conditions of original.

• Probably some effects non-existent or tiny and original was a fluke of chance and/or product of HARKing/p-hacking.
Easy to get 0% false alarm rate and 100% replication rate:

Just study tautologies!

Being vulnerable to error is integral to science.

But so too is error correction, and needs to improve error-correction practices.
There is evidence that many research psychologists use “questionable research practices” (QRPs) that undermine interpretation of $p$ values in null-hypothesis significance testing (NHST)

– Hypothesizing after the results are know (HARKing).
– Excluding subjects, observations, measures, or conditions
– Exploring potential moderators, transforms, alternative analyses
– Conducting many small studies but reporting only those that “worked”
How did these problems come about?

Powerful incentives to publish as many papers as possible.

Biggest rewards for solo/lead roles.

Emphasis on surprising findings (low prior odds).

Weak theories.

Measures that have low reliability and dubious sensitivity.

Poor understanding of fundamentals of NHST:
• Misunderstanding of what $p$ values mean
• Mistaken intuitions regarding statistical power
• Failure to appreciate how HARKing & $p$-hacking undermine $p$

Underestimation of risks of experimenter bias and exptr effects
How did these problems come about? – Part II

Direct replications discouraged, especially across labs.

Conceptual replications encouraged:

- Pashler (2018) speculated that failed conceptual replications are attributed to moderators and file-draweredy w/o reducing confidence in original effect, whereas successful conceptual reps are published and increase confidence in original effects.

Often difficult to get materials and data from authors.

When data obtained, often difficult to reproduce analyses.

These all undermine error-detection and correction.
How do we overcome these problems?

1. Foster awareness/understanding of statistical/methodological/theoretical issues.

2. Value quality over quantity, realness over surprisingness, accuracy over speed.

3. Promote transparency; emphasize disclosure not sales.

Let’s try to get more specific and concrete.
Steps you Can Take to Enhance Replicability


2. Consider transparency when applying to institutional review board (ethics committee) for permission to conduct study (e.g., get consent to post data, perhaps video sessions, etc.).

3. Before collecting data, record and preregister any plans for testing hypotheses (e.g., DVs, IVs, exclusions, transformations, moderators, etc.).
Steps you Can Take to Enhance Replicability

4. Before collecting data make plan for sample size.

5. Avoid focus on tiny effects, especially in between-subject designs.

6. Consider validity, reliability, and sensitivity of measures.


8. When possible ground project in strong theory.

9. If you obtain interesting results, consider attempting a direct replication.
Steps you Can Take to Enhance Replicability

10. Have plans/procedures for handling data files (how to name projects, how to name data files and variables, how to securely back up, how to de-identify).

11. Use a script-based analysis program rather than GUI for analyses (or if GUI then use “paste” to save syntax) so you have record of exactly how analysis was done.

12. If feasible, use the (securely backed-up) raw data file for all analyses, rather than creating filtered/transformed data files (which can get confusing to track).
Steps you Can Take to Enhance Replicability

13. Find a “data buddy” (Morey & Morey)

Attempt to reproduce each other’s analyses.

www.psychologicalscience.org/observer/habits-and-open-science

14. Report an index of precision (e.g., 95% confidence or credible intervals) around DV means estimated effect sizes.

15. Use fine-grained graphs showing distributions (scatter plots, box plots, pirate plots, frequency histograms).
Examine raw data before you analyze findings
Steps you Can Take to Enhance Replicability

16. Do not describe non-sig finding as evidence for null.

- Underpowered studies often fail to reject false nulls (Type II errors), sometimes with huge \( p \) values.
- If want to argue for null, consider Bayesian analyses (JASP is free and easy to use) or Equivalence Tests (also free -- http://shiny.ieis.tue.nl/TOST_vs_SGPV/)

Daniel Lakens

Consider Lakens’s online stats course, www.coursera.org/learn/statistical-inferences (and other such resources).
Steps you Can Take to Enhance Replicability

17. Don’t describe a pattern in which an effect is significant in one condition but not in another as an interaction.


19. Consider using Open Science Framework (OSF) etc. to organize/archive research projects.

20. Consider using the Psychological Science Accelerator to conduct multi-lab research.
21. Use tools such as StatCheck to detect errors in stats reporting.

**StatCheck**

Michele B. Nuijten
Tilberg University

Sacha Epskamp
Amsterdam University

Detects internal inconsistencies between elements of widely used inferential statistical tests (e.g., F, df, & p).
22. Teach students about these topics in methods, stats, and substantive courses. Many wonderful online tools to help students understand statistical issues (e.g., Felix Schönbrodtt’s [https://www.nicebread.de/](https://www.nicebread.de/) ; Daniel Lakens’s [http://daniellakens.blogspot.com/](http://daniellakens.blogspot.com/) and MOOC; Lorne Campbell’s course [http://www.lornecampbell.org/?p=171](http://www.lornecampbell.org/?p=171) ; Geoff Cumming’s [https://thenewstatistics.com/itns/](https://thenewstatistics.com/itns/) ).

23. As reviewer, attend to transparency/replicability issues (see in-press article w/ tips for reviewing for transparency-promotion [https://kar.kent.ac.uk/69135/](https://kar.kent.ac.uk/69135/) ).

24. Nominate people who are knowledgeable and care about these issues to positions in professional societies, departmental hiring committees, tenure committees, etc.
25. CONSIDER JOINING

http://improvingpsych.org/

Fourth annual meeting 7-9 July 2019 in Rotterdam.

Anita Eerland (chair)
Sara Weston (past chair)
Daniël Lakens
Rolf Zwaan
Psychology's Replication Renaissance