Leverhulme Lecture:

The Credibility Revolution in the Life Sciences

Slide Deck #2
Take-home points from ESCI

• High $p$ of detecting medium-sized effects in between-SS designs requires largish $N$.

• $p$ is noisy; a big $p$ doesn’t mean null is true; a small $p$ doesn’t promise replication.

• When power is too low, as here, a test cannot be sig unless the study yields a larger-than-average effect.
Publication Bias

Under standard NHST, non-significant results are considered ambiguous, not as evidence for the null.

Consequently, for decades it has been all but impossible to publish non-significant findings.

Hence many researchers conduct large numbers of under-powered studies and submit for publication the few that attain significance.
How/why did misuses of NHST come about? (Part II)

Powerful incentives to publish as many articles as possible.

Emphasis on surprising findings (low prior odds of an effect).

Biggest rewards for solo/lead roles (ergo competition, secrecy)

Weak/vague theories.

Measures that have low reliability and dubious validity.

Poor understanding of fundamentals of NHST:
• Misunderstanding of what \( p \) values mean
• Mistaken intuitions regarding statistical power
• Selective publication of studies with statistically significant results
• Failure to appreciate how HARKing & \( p \)-hacking undermine \( p \)
HARKing

Norbert L. Kerr (1989): Hypothesizing After the Results are Known

Explore rich data set for patterns, then use NHST to decide if a pattern is “significant.”

Statistician Andrew Gelman calls this “the garden of forking paths.”

The tests conducted are inspired by the data you happened to obtain, so $p$ values cannot be interpreted at face value.
**False-positive Psychology**

Joseph Simmons, Leif Nelson, and Uri Simonsohn (2011)

**P Hacking: Exploiting researchers’ degrees of freedom in ways that exaggerate effect size:**
- Not reporting all of the measures and/or conditions tested.
- Dropping subjects/observations to attain $p < .05$.
- Exploring co-variates or data transforms.
- “Optional stopping” in data collection.

Combining several p-hacks can hugely inflate false discovery rate.
Experimenter effects = Research participants may change behavior/reports in line with researcher’s hypotheses.

Recent evidence that experimenters can influence research participants’ responses on computer-run tasks (Gilder & Heerey, 2018).

Thandiwe Gilder

Erin Heerey
Summary of the Bad News

Incentives + low power + $p$ hacking + experimenter effects + publication bias contribute to effect-size exaggeration in published literature.

Direct replications not encouraged in many life sciences.

Often difficult to get materials and data from authors. If data are obtained, often difficult to reproduce analyses due to ambiguities and/or errors.

These all undermine error-detection and correction.

But now for the good news!
Support, enable, and reward

- Competence
- Transparency
- Replicability/reproducibility
- Teamwork
- Modesty/caution

Workshop on Tools for Transparency
Tuesday 15 October 3-5pm
Anglesea Bldg 0.28

THE CREDIBILITY REVOLUTION IN LIFE SCIENCES
Brian Nosek, co-founder of the Center for Open Science
Preregistration of research plans:

Before looking at data, make an immutable record of your plans (e.g., sample size, predictions, criteria for testing predictions).

Can always deviate from plan. Just not blindly or secretly.
Tools to Support Crafting of High-Quality Preregistrations

Aczel et al. (in press). Transparency Checklist for Social and Behavioural Researchers
*Nature Human Behaviour*

[Link to shiny app version of checklist!](#)

Balazs Aczel

[The Center for Open Science’s preregistration checklist.](#)
Badges signal articles that provide quick and easy access to data or materials or that were preregistered.
Percentage of *Psych Science* Articles With Badges

*Jan-Sep 2019*
Registered Reports: Peer Review prior to (as well as after) data collection.

Publication decision based on perceived importance and soundness of the proposed research, not on the results.

Chris Chambers
Cardiff Uni
✓ New norms and technology for developing, documenting, and maintaining an orderly workflow, safeguarding records, standardizing lab routines, etc.

✓ New norms and support for ethical data sharing **Practical Tips for Ethical Data Sharing**

✓ Tools for learning/teaching **statistical concepts**

✓ Tools for sample-size planning by **Westfall** and by **Lakens**

✓ Free software for making deeply informative data graphs.
Bad Old Graph
From https://rpubs.com/crazyhottommy/reorder-boxplot
✓ New norms and technology for developing, documenting, and maintaining an orderly workflow, safeguarding records, standardizing lab routines, etc.

✓ New norms and support for ethical data sharing Practical Tips for Ethical Data Sharing

✓ Tools for learning/teaching statistical concepts

✓ Tools for sample-size planning by Westfall and by Lakens

✓ Free software for making deeply informative data graphs.

✓ Tools for assessing support for null hypothesis (Bayesian analyses, equivalence tests)

✓ Free online stats courses such as Daniel Lakens’s Coursera (which I hope to take soon!)

✓ New spirit of cooperative striving together to advance science
The Psychological Science Accelerator

A globally distributed network of psychological science laboratories (currently over 500), representing over 70 countries on all six populated continents, that coordinates data collection for democratically selected studies.
Life Science’s Replication Renaissance