ARE RESEARCH FINDINGS RELIABLE?
SAINT KIZITO OMALA | MAKERERE UNIVERSITY

2016 EAST AFRICA IMPACT EVALUATION WORKSHOP AND EVIDENCE SUMMIT
When you read or hear about a new piece of research (i.e. in a publication or a talk), what are things that would make you trust the result? What would make you skeptical?

What problems undermine our ability to rely on scientific research?
RELIABLE RESEARCH

- P-hacking
- Selective reporting
- Failure to replicate
- Lack of transparency
- Publication bias
P-HACKING:

• Researchers test hypotheses i.e. look for relationships among variables (e.g. schooling, test scores).
  • In particular, a result which shows a result which is statistically significant at $p<.05$ is often considered noteworthy and thus more publishable.
• This leads to pressure to tinker with specifications (how variables are coded, which outliers are excluded, which subgroups are considered, etc) to find $p<.05$. 
P-HACKING:

- “Why Most Research Findings Are False” Ioannidis (PLOS 2005)
  - “The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.”
- This flexibility, also known as “researchers degrees of freedom,” can undermine the reliability of published results and can give way to conscious and unconscious p-hacking
P-HACKING

- Also called “data fishing”, “data mining”, “data massaging”.
- “If you torture your data long enough, they will confess to anything.” – Ronald Coase

“If you don’t reveal some insights soon, I’m going to be forced to slice, dice, and drill!”
SELECTIVE REPORTING:

- Cherry picking results for reporting
PUBLICATION BIAS

"File drawer problem"
(Rosenthal Robert, 1979)
PUBLICATION BIAS:

- Statistically significant results more likely to be published, while null results are buried in the file drawer
VERY FEW STUDIES REPORT NULL RESULTS:

**ACCENTUATE THE POSITIVE**

A literature analysis across disciplines reveals a tendency to publish only 'positive' studies — those that support the tested hypothesis. Psychiatry and psychology are the worst offenders.

- Physical
  - Space sciences
  - Geosciences
  - Environment/Ecology
- Biological
  - Plant and animal sciences
  - Computer science
  - Physics
  - Neuroscience and behaviour
  - Microbiology
  - Chemistry
  - Social sciences
  - Immunology
  - Molecular biology and genetics
  - Economics and business
  - Biology and biochemistry
  - Clinical medicine
  - Pharmacology and toxicology
  - Materials science
  - Psychiatry/psychology

Proportion of papers supporting tested hypothesis

50% 60% 70% 80% 90%
IN SOCIAL SCIENCES:

Most null results are never written up
The fate of 221 social science experiments

Source: A. Franco et al., Science (28 August)
IN ECONOMICS...

IN POLITICAL SCIENCE

Gerber and Malhotra, 2008
doi: 10.1561/1000.00008024

Figure 1(a). Histogram of z-statistics, *APSR* & *AfPS* (Two-Tailed).
POP QUIZ:

• How would you respond to this person?
  • “I don’t see the problem if we don’t see the null results. We’re interested in finding out what works, and reading about studies where the researchers didn’t find anything significant to report won’t help us much!”
  • If researchers degrees of freedom undermines the reliability of findings and p-hacking gives way to exaggerated results, keeping null results in the file drawer can bias research and have real world implications. The null results help to provide the full picture.
“Replication” is often used to mean different things. Here are three different activities which replication refers to:

(1) **Verification and re-analysis**: Checking that original data/code can produce published results, as well as going further to check the robustness of the results.

(2) **Reproduction**: Testing whether the results hold up when the study conducted in a very similar way.

(3) **Extension**: Investigating whether the results hold up when the study is conducted in another place, under different conditions, etc., to test external validity.
REPLICATIONS:

- Specification (covariates, treatment of outliers, etc)
  - Different
  - Same

- Population
  - Different
  - Same

- Re-analysis
  - New data: Reproduction
  - Old data: Verification

- Re-analysis and extension
  - Extension
WELL-KNOWN REPLICATIONS:

• Reinhart-Rogoff (2013)
  • Spreadsheet errors found by grad student (Herndon et al 2013)
• Deworming debate (2015)
  • Miguel/Kremer (2004) and Aiken/Davey et al. (2015)
  • Big debate within development econ/epidemiology
• Reproducibility project in psychology (2015)
  • ~40% of studies successfully reproduced.
  • What does it mean for a replication to “fail”? 
REPLICATION INCENTIVES:

• Journals are often reluctant to publish “unoriginal” work, and replications are often considered “unoriginal.”

• Yet simply sharing results of replications on a website (e.g. Political Science Replication Initiative; Psych file drawer) is not ideal either: some argue strongly that replications should be peer-reviewed
TRANSPARENCY:

Peng 2011, Science 334(6060) pp. 1226-1227
• Sharing data, code and survey instruments (along with clear documentation) can allow others to check one’s work
Yet relatively few researchers sharing data:

- **Alsheick-Ali et al. 2011**: Review of 10 first research papers of 2009 published in top 50 journals by impact factor. Of the 500 papers, 251 were subject to a data availability policy of some kind: 59% did not adhere to the policy, most commonly by not publicly depositing the data (73%). Overall, only 47 papers (9%) deposited the full primary raw data online.

- Research funders (e.g. government funders such as NSF, NIH, ERSC in the UK, private foundations) are adopting data-sharing policies at greater rates. Yet they rarely enforce their policies.
And relatively few journals have data-sharing policies:

- In 2013, only 18 out of 120 political science journals inspected had data-sharing policies (Gherghinaa and Katsanidoua 2013) and another review found that only 29 out of 141 journals reviewed in economics had policies (Vlaeminck 2013).

- However, having a mandatory policy makes it much more likely that researchers will share data:
  - The rate of data archiving compliance for surveyed journals with the strictest policies, which required data archiving along with data accessibility statements in manuscripts, was nearly 1000x higher than having no policy (Vines 2014).