13th Munin Conference on Scholarly Publishing, Tromsø, November 2018

The replication and reproducibility crises: origins and consequences for studies of ecology and evolution

Nigel G. YOCCOZ
Sandra HAMEL
UiT The Arctic University of Norway
Tromsø

Outline

- The origins of the «replication» crises
- Many definitions and aspects of replication, reproducibility, robustness etc.
- Some of the factors contributing to the crises
- Well-documented in fields like psychology, what about ecology and evolution ?
- Transferability as another way to look at replication
- Some ways forward – reforming education and scholarly publishing

RESEARCH ARTICLE SUMMARY

SCIENCE

28 AUGUST 2015 • VOL 349 ISSUE 6251

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

The mean effect size of the replication effects was half the magnitude of the mean effect size of the original effects, representing a substantial decline.

97 percent of original studies had significant results (P < .05).
36 percent of replications had significant results

The Economist 2013

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

The New York Times

SCIENCE TIMES AT 40

Essay: The Experiments Are Fascinating. But Nobody Can Repeat Them.

Science is mired in a “replication” crisis. Fixing it will not be easy.

By Andrew Gelman

Nov. 19, 2018

SCIENCE

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

ED YONG NOV 19, 2018

Many Labs 2

14 out of 28 cases

Ironically enough, it seems that one of the most reliable findings in psychology is that only half of psychological studies can be successfully repeated.

Main criticisms of replication studies:

1) the replication attempts themselves might be too small.
2) the researchers involved might be incompetent, or lack the know-how to properly pull off the original experiments.
3) people vary, and two groups of scientists might end up with very different results if they do the same experiment on two different groups of volunteers.

(Yong 2018 The Atlantic)

Western, Educated, Industrialized, Rich and Democratic (WEIRD)

INTRODUCTION

Again, and Again, and Again …

SCIENCE VOL 334 2 DECEMBER 2011

REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard.

Data Replication & Reproducibility

CONTENTS

Perspectives

SCIENTIFIC INTEGRITY

What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of "research reproducibility" are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for "truth."

methods reproducibility
results reproducibility
inferential reproducibility

Reproducibility vs. Replicability: A Brief History of a Confused Terminology

Hans E. Plesser[1,2*]

TABLE 1 | Comparison of terminologies. See text for details.

Goodman	Claerbout	ACM
		Repeatability
Methods reproducibility	Reproducibility	Replicability
Results reproducibility	Replicability	Reproducibility
Inferential reproducibility		

What factors contribute to the crisis?

Misunderstanding and/or misuse of statistical methods contribute quite «significantly» (P<0.05)

See poster by Sandra Hamel

Table 2. Terminology to describe practices that introduce or hide multiplicity.
Multiple comparisons (many statisticians)
File-drawer problem (*29*)
Pseudoreplication (*32*)
Significance questing (*33*)
Data mining, dredging, torturing (*34*)
Hypothesizing after the results are known (HARKing) (*30*)
Data snooping (*35*)
Selective outcome reporting (*36*)
Silent multiplicity (*37*)
Specification searching (*38*)
P-hacking (*31*)

SCIENTIFIC INTEGRITY

What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

American Scientist, Volume 102

Would the same data-analysis decisions have been made with a different data set?

PLOS | ONE

RESEARCH ARTICLE

Questionable research practices in ecology and evolution

Hannah Fraser[1]*, Tim Parker[2], Shinichi Nakagawa[3], Ashley Barnett[1], Fiona Fidler[1,4]

At least half of evolutionary biologists and ecologists fudge results, survey finds

Bio science faces a "replication crisis" as big as the one currently blighting psychology. Andrew Masterson reports.

Across the two groups, ***we found 64% of surveyed researchers reported they had at least once failed to report results because they were not statistically significant*** (cherry picking);
42% had collected more data after inspecting whether results were statistically significant (a form of *p* hacking) and

51% had reported an unexpected finding as though it had been hypothesised from the start (HARKing).

Such practices have been directly implicated in the low rates of reproducible results uncovered by recent large scale replication studies in psychology and other disciplines.

The rates of QRPs found in this study are comparable with the rates seen in psychology, indicating that the reproducibility problems discovered in psychology are also likely to be present in ecology and evolution.

(Fraser et al. 2018 Plos One)

Informative Irreproducibility and the Use of Experiments in Ecology

BioScience • *October 2018 / Vol. 68 No. 10*

ANTHONY R. IVES

A key distinction, often not made, is between ***reproducibility among experiments conducted at different times, on different systems***, or with different methods, and ***reproducibility within the same experiment*** that could be achieved by increasing sample size.

Experiments should be judged on what they tell us about the system under study in a ***strict statistical way***. And they should be judged on whether they are ecologically interesting, giving information that provides ***qualitative insights into other systems***.

But they ***should not be judged on whether they can be reproduced to allow quantitative statistical comparisons among experiments if this is not their intended design.***

Indeed, the agreed wisdom in implementation science is that context effects in healthcare are so profound, that we should actually expect to see variations in outcome every time we repeat an intervention in a new setting.

In other words, this received wisdom suggests that by definition, differences in research outcome should be ascribed to changes in context, rather than a failure to replicate an earlier study.

Does health informatics have a replication crisis?

Enrico Coiera,[1] Elske Ammenwerth,[2] Andrew Georgiou,[1] and Farah Magrabi[1]

Metaresearch for Evaluating Reproducibility in Ecology and Evolution

FIONA FIDLER, YUNG EN CHEE, BONNIE C. WINTLE, MARK A. BURGMAN, MICHAEL A. MCCARTHY AND ASCELIN GORDON

We argue that these conditions constitute sufficient reason to systematically evaluate the reproducibility of the evidence base in ecology and evolution. In some cases, the direct replication of ecological research is difficult because of strong temporal and spatial dependencies, so here, we propose metaresearch projects that will provide proxy measures of reproducibility

Transferability is a form of inferential reproducibility (spatial and temporal)

Trends in Ecology & Evolution

CellPress REVIEWS

Review

Outstanding Challenges in the Transferability of Ecological Models

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Transferability of biotic interactions: Temporal consistency of arctic plant–rodent relationships is poor

Eeva M. Soininen[1] | John-Andre Henden[1] | Virve T. Ravolainen[2] | Nigel G. Yoccoz[1] | Kari Anne Bråthen[1] | Siw T. Killengreen[1] | Rolf A. Ims[1]

Seasonal difference in temporal transferability of an ecological model: near-term predictions of lemming outbreak abundances

Eivind Flittie Kleiven, John-André Henden, Rolf Anker Ims & Nigel Gilles Yoccoz

Ecological models have been criticized for a lack of validation of their temporal transferability.

SCIENTIFIC REPORTS | (2018) 8:15252

Conclusion: This has been emphasized before…

Fisher (1934) “Statistical Methods for Research Workers”

Page 3: ‘the salutary habit of repeating important experiments, or of carrying out original observations in replicate’.

Page 123: ‘confidence to be placed in a result depends not only on the magnitude of the mean value obtained, but equally on the agreement between parallel experiments’.

Change in statistical education and reviewing process:

The replication crisis in science is often presented as an issue of scientific procedure or integrity. But all the careful procedure and all the honesty in the world won’t help if your signal (the pattern you’re looking for) is small, and the variation (all the confounders, the other things that might explain this pattern) is high.

The big problem in science is not cheaters or opportunists, but sincere researchers who have unfortunately been trained to think that every statistically “significant” result is notable.

(Gelman 2018)

Change in what is seen as «important research»

Journal of Animal Ecology 2013, **82**, 1–2 doi: 10.1111/1365-2656.12026

EDITORIAL

Publishing the best original research in animal ecology: looking forward from 2013

Tim Coulson[1]*, Graeme Hays[2], Mike Boots[3], Ken Wilson[4], Liz Baker[5] and Peter Livermore[5]

The journal aims to publish only ***novel and exciting papers*** that will appeal to researchers across the field of animal ecology; that is, by advancing current ecological theory and generating insights that ***extend beyond the study system in question***. While we appreciate the value of papers describing aspects of the ecology or life-history of a single species or that ***verify previously known insights in a new system***, we believe that ***these are more likely to reach their target audience if published in the more specialized literature***

Beyond ensuring "correctness," the goal of these efforts, and I would argue their primary goal, ***should be to enable future scientists to build upon the work to go further***.

Before attributing difficulties with reproducibility, replicability, robustness, and generalizability to a dim view of our fellow scientists as being sloppy, biased, or untrustworthy, ***it is worth seriously considering the many factors— biological, statistical, and sociological— that pose a threat***.

Although there is much room for improvement, we must acknowledge that science is ***a process of learning and that it is really freaking hard.***

Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research

Patrick D. Schloss[a]

AMERICAN SOCIETY FOR MICROBIOLOGY mBio®

May/June 2018 Volume 9 Issue 3 e00525-18