Reproducibility, replication, etc.

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

February 22, 2019

Questions

How should one evaluate research results?

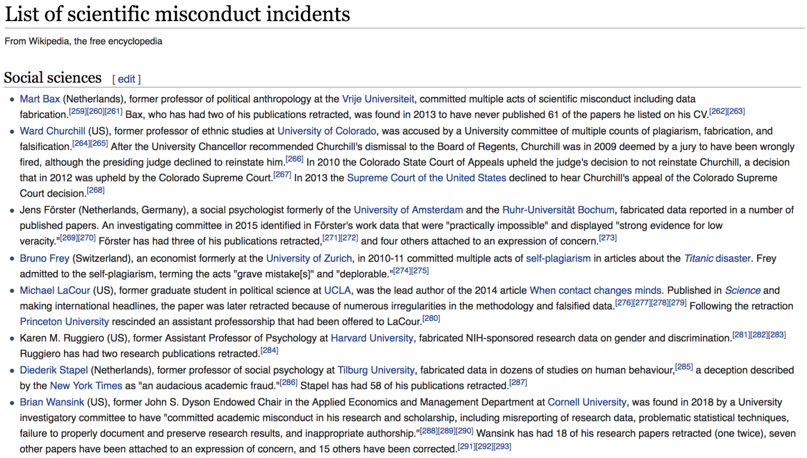
- Was the research done and reported honestly / correctly?
- Is the result “real” or an artifact of the data / analysis?
- Will it hold up over time?
- How robust is the result to small changes?
- How important / useful is the finding?

Honesty

Was the data accurately collected and reported?

We'll take the optimistic view that most researchers are honest, although there are exceptions

Honesty



List of scientific misconduct incidents

From Wikipedia, the free encyclopedia

Social sciences [edit]

- Mart Bax (Netherlands), former professor of political anthropology at the Vrije Universiteit, committed multiple acts of scientific misconduct including data fabrication.[259][260][261] Bax, who has had two of his publications retracted, was found in 2013 to have never published 61 of the papers he listed on his CV.[262][263]
- Ward Churchill (US), former professor of ethnic studies at University of Colorado, was accused by a University committee of multiple counts of plagiarism, fabrication, and falsification.[264][265] After the University Chancellor recommended Churchill's dismissal to the Board of Regents, Churchill was in 2009 deemed by a jury to have been wrongly fired, although the presiding judge declined to reinstate him.[266] In 2010 the Colorado State Court of Appeals upheld the judge's decision to not reinstate Churchill, a decision that in 2012 was upheld by the Colorado Supreme Court.[267] In 2013 the Supreme Court of the United States declined to hear Churchill's appeal of the Colorado Supreme Court decision.[268]
- Jens Förster (Netherlands, Germany), a social psychologist formerly of the University of Amsterdam and the Ruhr-Universität Bochum, fabricated data reported in a number of published papers. An investigating committee in 2015 identified in Förster's work data that were "practically impossible" and displayed "strong evidence for low veracity."[269][270] Förster has had three of his publications retracted,[271][272] and four others attached to an expression of concern.[273]
- Bruno Frey (Switzerland), an economist formerly at the University of Zurich, in 2010-11 committed multiple acts of self-plagiarism in articles about the *Titanic* disaster. Frey admitted to the self-plagiarism, terming the acts "grave mistake[s]" and "deplorable."[274][275]
- Michael LaCour (US), former graduate student in political science at UCLA, was the lead author of the 2014 article When contact changes minds. Published in *Science* and making international headlines, the paper was later retracted because of numerous irregularities in the methodology and falsified data.[276][277][278][279] Following the retraction Princeton University rescinded an assistant professorship that had been offered to LaCour.[280]
- Karen M. Ruggiero (US), former Assistant Professor of Psychology at Harvard University, fabricated NIH-sponsored research data on gender and discrimination.[281][282][283] Ruggiero has had two research publications retracted.[284]
- Diederik Stapel (Netherlands), former professor of social psychology at Tilburg University, fabricated data in dozens of studies on human behaviour,[285] a deception described by the New York Times as "an audacious academic fraud."[286] Stapel has had 58 of his publications retracted.[287]
- Brian Wansink (US), former John S. Dyson Endowed Chair in the Applied Economics and Management Department at Cornell University, was found in 2018 by a University investigatory committee to have "committed academic misconduct in his research and scholarship, including misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship."[288][289][290] Wansink has had 18 of his research papers retracted (one twice), seven other papers have been attached to an expression of concern, and 15 others have been corrected.[291][292][293]

Reproducibility

Can you independently verify the exact results using the *same data* and the *same analysis*?

Though a low bar, most research doesn't currently pass this test:

- Data or code aren't available / complete
- Code is difficult to run / understand
- Complex software dependencies

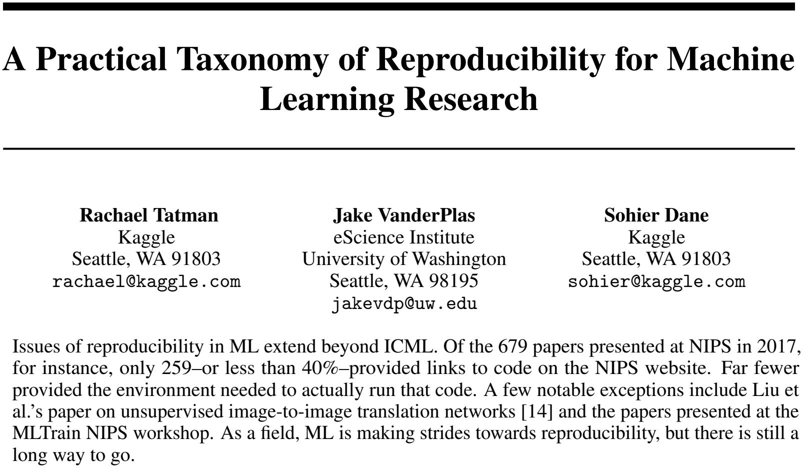
Reproducibility

Can you independently verify the exact results using the *same data* and the *same analysis*?

This is improving with better software engineering practices among researchers:

- Literate programming (Jupyter, Rmarkdown)
- Automated build scripts (Makefiles)
- Containers (Docker, Code Ocean)

Reproducibility



A Practical Taxonomy of Reproducibility for Machine Learning Research

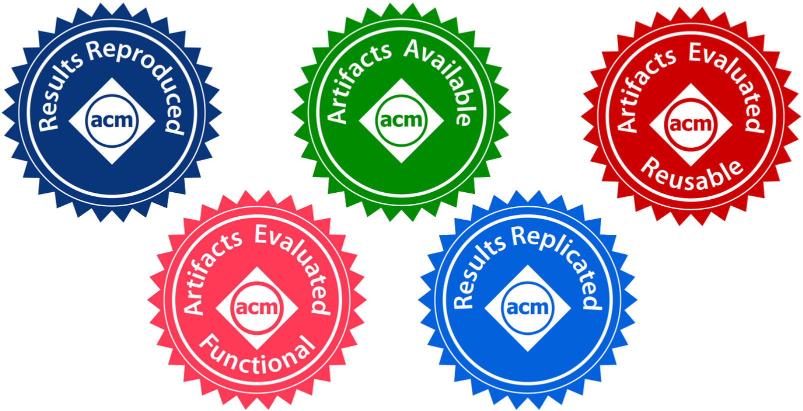
Rachael Tatman
Kaggle
Seattle, WA 91803
rachael@kaggle.com

Jake VanderPlas
eScience Institute
University of Washington
Seattle, WA 98195
jakevdp@uw.edu

Sohier Dane
Kaggle
Seattle, WA 91803
sohier@kaggle.com

Issues of reproducibility in ML extend beyond ICML. Of the 679 papers presented at NIPS in 2017, for instance, only 259–or less than 40%–provided links to code on the NIPS website. Far fewer provided the environment needed to actually run that code. A few notable exceptions include Liu et al.'s paper on unsupervised image-to-image translation networks [14] and the papers presented at the MLTrain NIPS workshop. As a field, ML is making strides towards reproducibility, but there is still a long way to go.

Reproducibility

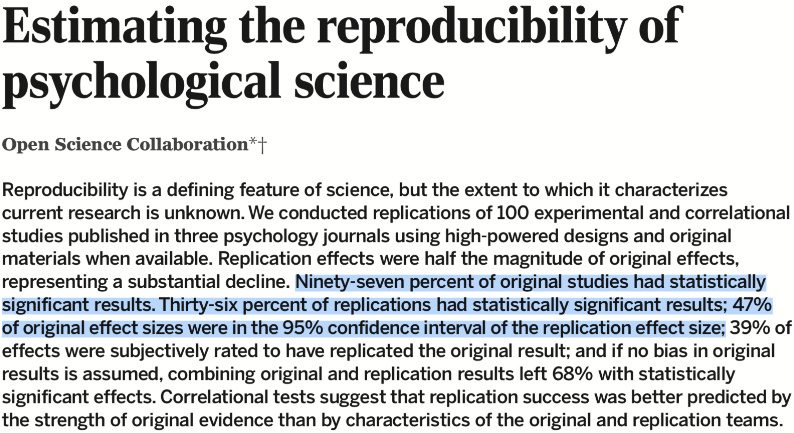


Replicability

Will the result hold up with new data but the same analysis?

- It’s easy to be fooled by randomness
- Noise can dominate signal in small datasets
- Asking too many questions of the data can lead to overfitting

Crisis



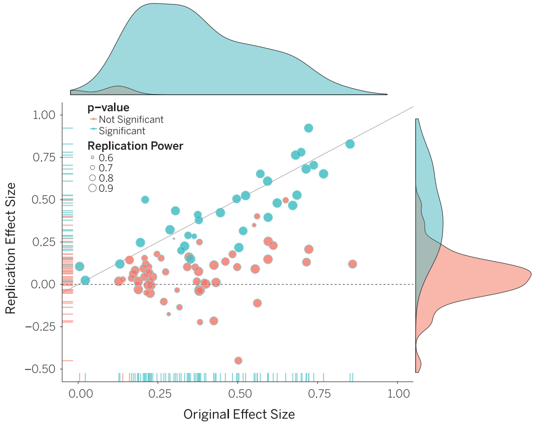
Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Believe about half of what you read

Crisis



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Believe about half of what you read

Statistics!

Hypothesis testing?
P-values?
Statistical significance?
Confidence intervals??
Effect sizes???

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

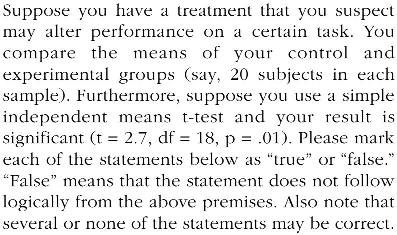
xkcd.com/892

Quiz #1 (h/t Shane Frederick)

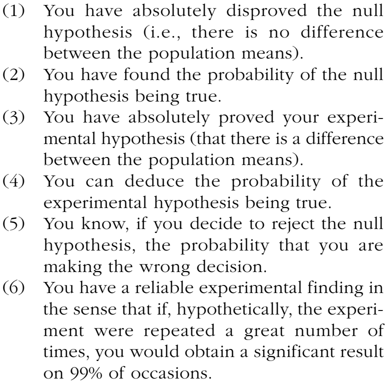
Which treatment would you prefer?

- Treatment A was found to improve health over a placebo by 10 points on average (with a standard error of 5 points) in a study with N = 100 participants.
- Treatment B was found to improve health over a placebo by 10 points on average (with a standard error of 5 points) in a study with N = 1,000 participants.

Quiz #2 (Oakes 1986)



Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant (t = 2.7, df = 18, p = .01). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.



(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).

(2) You have found the probability of the null hypothesis being true.

(3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

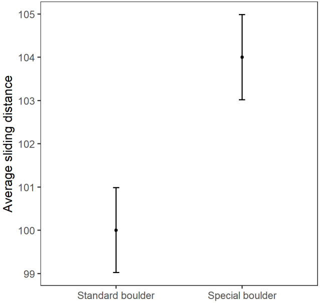
(4) You can deduce the probability of the experimental hypothesis being true.

(5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.

(6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Quiz #3a (Hofman, Hullman, Goldstein 2019?)

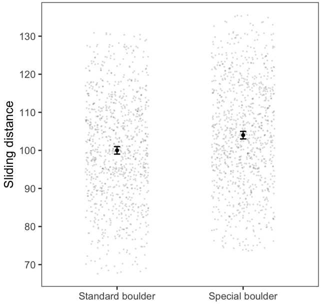
Below are results of an experiment with 1,000 slides of a standard boulder (left) and special boulder (right), with bars showing two standard errors on the mean.



Estimate the probability that a slide of the special boulder goes farther than a slide of the standard boulder.

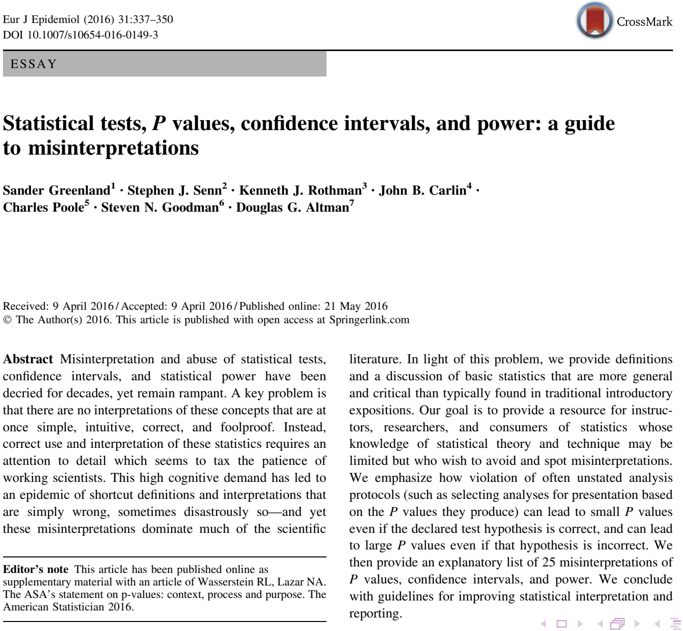
Quiz #3b (Hofman, Hullman, Goldstein 2019?)

Below are results of an experiment with 1,000 slides of a standard boulder (left) and special boulder (right), with bars showing two standard errors on the mean and points showing individual slides.



Estimate the probability that a slide of the special boulder goes farther than a slide of the standard boulder.

Misunderstandings



Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

CrossMark

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

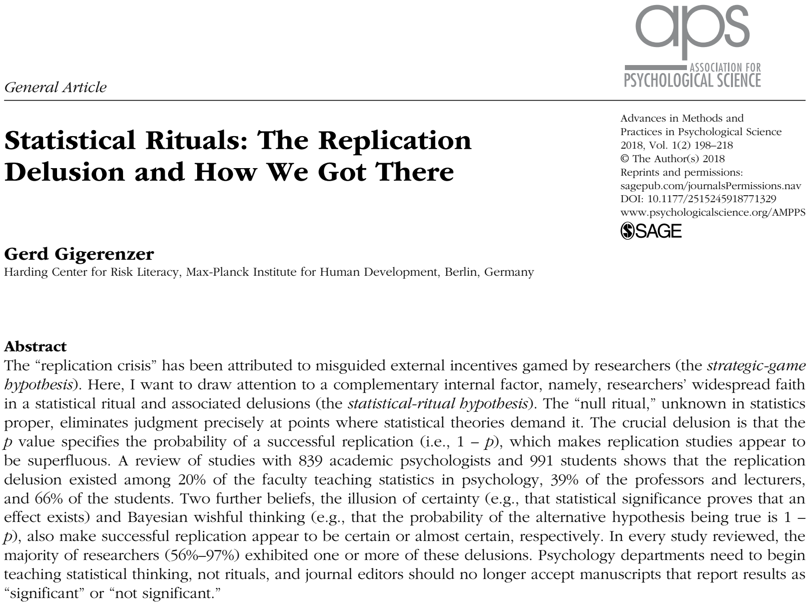
Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

Received: 9 April 2016 / Accepted: 9 April 2016 / Published online: 21 May 2016

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process and purpose. The American Statistician 2016.

Statistical rituals



General Article

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(2) 198–218
© The Author(s) 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2515245918771329
www.psychologicalscience.org/AMPPS

Statistical Rituals: The Replication Delusion and How We Got There

Gerd Gigerenzer
Harding Center for Risk Literacy, Max-Planck Institute for Human Development, Berlin, Germany

Abstract

The “replication crisis” has been attributed to misguided external incentives gamed by researchers (the *strategic-game hypothesis*). Here, I want to draw attention to a complementary internal factor, namely, researchers’ widespread faith in a statistical ritual and associated delusions (the *statistical-ritual hypothesis*). The “null ritual,” unknown in statistics proper, eliminates judgment precisely at points where statistical theories demand it. The crucial delusion is that the p value specifies the probability of a successful replication (i.e., $1 - p$), which makes replication studies appear to be superfluous. A review of studies with 839 academic psychologists and 991 students shows that the replication delusion existed among 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students. Two further beliefs, the illusion of certainty (e.g., that statistical significance proves that an effect exists) and Bayesian wishful thinking (e.g., that the probability of the alternative hypothesis being true is $1 - p$), also make successful replication appear to be certain or almost certain, respectively. In every study reviewed, the majority of researchers (56%–97%) exhibited one or more of these delusions. Psychology departments need to begin teaching statistical thinking, not rituals, and journal editors should no longer accept manuscripts that report results as “significant” or “not significant.”