Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Psychology in Crisis

Overview

- Crisis? What Crisis?
  - Scientific Misconduct
  - Confirmation Bias
  - False Positive Psychology
  - Questionable Research Practices
  - Multiple Comparisons
  - Exploratory vs. Confirmatory Research
- Approaches to a Solution
  - Preregistered Replications
- Reproducibility in Psychology
  - Systematic Replication of Psychological Experiments
- Light at the End of the Tunnel . . .

Crisis? What Crisis?

Scientific misconduct

- Diederik Stapel (former professor of social psychology at Tilburg University)
  - Stapel fabricated and manipulated data for his publications
  - At least 38 articles, published in the most important journals of social psychology, are based on fabricated data
  - As of 2019, Stapel has had 58 of his publications retracted
- Jens Förster (former professor of social psychology at University of Amsterdam and Ruhr-Universität Bochum)
  - Allegations concerning the manipulation of data based on statistical anomalies in several publications
  - Withdrawal of a Humboldt professorship
  - According to a settlement with the court of honor of the German Psychological Society (DGPs) two articles were retracted
  - Jens Förster always refuted the charges

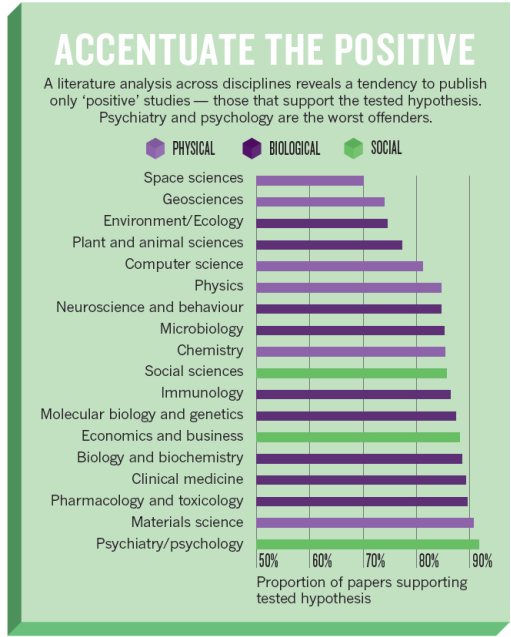
Scientific Misconduct

Problems

- Fabrication of data is possible
- Data manipulations remained undetected over decades
- Practically no direct replications of published findings
- Psychological journals only published articles considered original research
- So-called "conceptual replications" do not admit any conclusions on the original findings
  - Operationalizations of the phenomenon, independent and dependent variables, type and design of the study, or participants may differ substantially

Confirmation Bias

Yong, E. (2012). Bad copy. Nature, 485, 298-300.



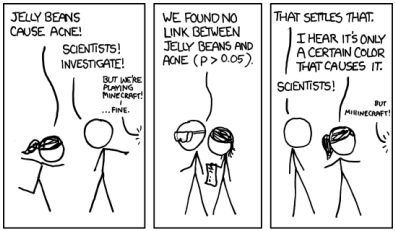
Confirmation Bias

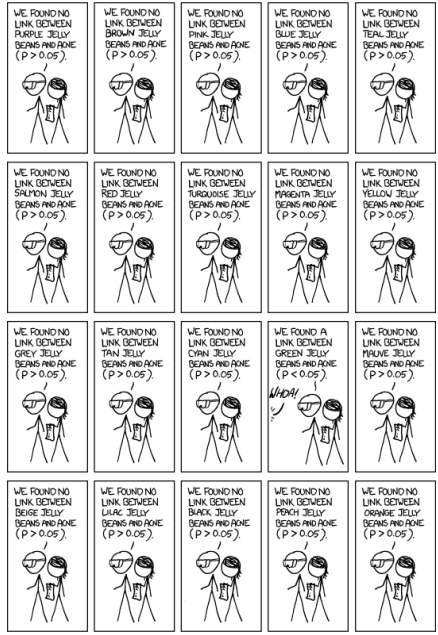
In Psychology more than 90% of the studies actually find the postulated effects

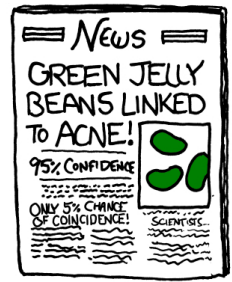
Possible reasons

- Studies reporting a null effect are often not accepted for publication in psychology journals
- This causes a systematic selection bias (file drawer effect) and overestimates the actual effect sizes
- In such an environment researchers may be tempted
  - to conduct several studies and to report only those that produced significant results (selective reporting)
  - to use methods that tend to produce significant results (questionable research practices, *p*-hacking, ...)

Selective Reporting







Cartoon from xkcd by Randall Munroe, `https://xkcd.com/882/`

False Positive Psychology

- Identifying scientific significance with statistical significance (e. g., a p-value $p < 0.05$) leads to publishing false-positive results
- The seminal paper

  Ioannidis, J.P.A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. `https://doi.org/10.1371/journal.pmed.0020124`

  considers the probability of observing an effect given a significant result in the context of a Bayesian approach
- This framework can also handle other systematic factors (such as publication bias)

False Positive Psychology

Positive predictive value $P(\text{effect} \mid \text{significant})$

- Given

$$\alpha = P(\text{significant} \mid \text{no effect})$$
$$\beta = P(\text{not significant} \mid \text{effect})$$
$$R = \frac{P(\text{effect})}{P(\text{no effect})}$$

one computes

$$P(\text{effect} \mid \text{sign.}) = \frac{P(\text{sign.} \mid \text{effect}) \cdot P(\text{effect})}{P(\text{sign.})}$$
$$= \frac{P(\text{sign.} \mid \text{effect}) \cdot P(\text{effect})}{P(\text{sign.} \mid \text{effect}) \cdot P(\text{effect}) + P(\text{sign.} \mid \text{no eff.}) \cdot P(\text{no eff.})}$$

False Positive Psychology

Positive predictive value

- This leads to

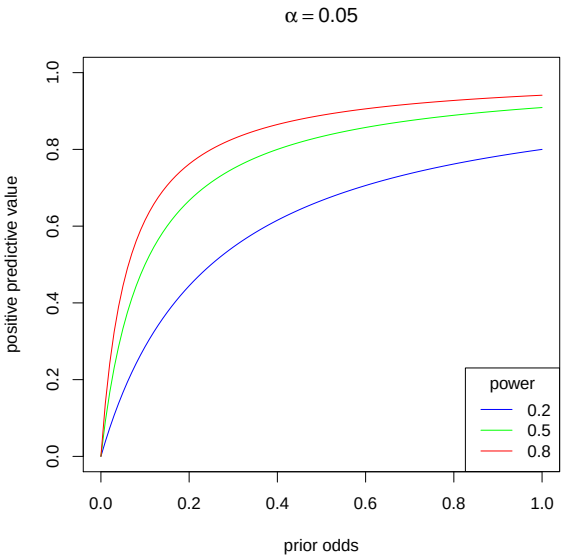
$$P(\text{effect} \mid \text{significant}) = \frac{(1-\beta)\cdot R}{(1-\beta)\cdot R + \alpha}$$

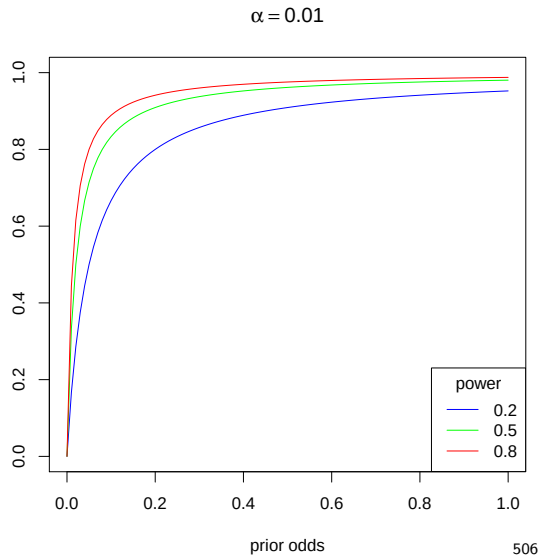
  with $1-\beta$ the power of the statistical test and R the prior odds for the existence of the effect

- The following diagrams assume that the nominal α- and β-levels actually hold true (which often may not be the case, see below)

False Positive Psychology

https://shinyapps.org/apps/PPV/





Questionable Research Practices

The often-quoted paper

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366.
`https://doi.org/10.1177/0956797611417632`

considers the degrees of freedom in data analysis

- Should more data be collected?
- Should some observations be excluded?
- Which conditions should be combined and which ones compared?
- Which control variables should be considered?
- Should specific measures be combined or transformed or both?

Questionable Research Practices

- This exploratory behavior is not the by-product of malicious intent, but rather the result of two factors
  - Ambiguity in how best to make these decisions
  - The researcher's desire to find a statistically significant result
- Simulations were run to assessed the impact of four common degrees of freedom
  - A Flexibility in choosing among dependent variables
  - B Flexibility in choosing sample size
  - C Flexibility in using covariates
  - D Flexibility in reporting subsets of experimental conditions

Questionable Research Practices

- Simulation results (Simmons et al., 2011)

  The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three t tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one t test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a t test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender $\times$ Condition interaction was significant. Results for Situation D were obtained by conducting t tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = –1, medium = 0, high = 1).

Questionable Research Practices

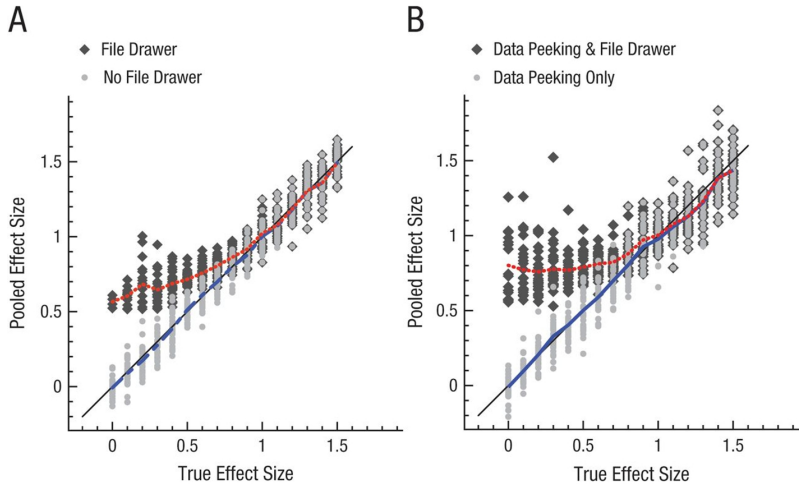
- Simulation results (Simmons et al., 2011)

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Questionable Research Practices

- Simulation results (Francis, 2012)



Francis, G. (2012). The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7, 580-589.

Multiple Comparisons

- Andrew Gelman points out in various publications as well as on his blog

  `https://statmodeling.stat.columbia.edu/`

  that data analysis is usually based on a multitude of decisions ("garden of forking paths")
- These (data dependent) decisions put the researcher in a situation of multiple comparisons (leading to Type I error inflation)
- This also holds in case of performing only a single analysis with no "*p*-hacking" (cf. Simmons et al., 2011)
- Gelman & Loken (2013) elaborate on this issue in some exemplary experiments

Multiple Comparisons

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡

14 Nov 2013

"I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world." — Borges (1941)

Abstract

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

`http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`

Exploratory vs. Confirmatory Research

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J. & Kievit, R.A. (2012). An agenda for purely confirmatory research. Perspectives on Psychological Science, 7, 632-638. https://doi.org/10.1177/1745691612463078



Fig. 1. A continuum of experimental exploration and the corresponding continuum of statistical wonkiness. On the far left of the continuum, researchers find their hypothesis in the data by post-hoc theorizing, and the corresponding statistics are "wonky," dramatically overestimating the evidence for the hypothesis. On the far right of the continuum, researchers preregister their studies such that data collection and data analyses leave no room whatsoever for exploration, and the corresponding statistics are "sound" in the sense that they are used for their intended purpose. Much empirical research operates somewhere in between these two extremes, although for any specific study the exact location may be impossible to determine. In the grey area of exploration, data are tortured to some extent, and the corresponding statistics are somewhat wonky. Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

Approaches to a Solution

Recommendations (Simmons et al., 2011)

Table 2. Simple Solution to the Problem of False-Positive Publications

Requirements for authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Guidelines for reviewers

1. Reviewers should ensure that authors follow the requirements.
2. Reviewers should be more tolerant of imperfections in results.
3. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
4. If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.

Approaches to a Solution

Recommendations (Simmons et al., 2011)

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~20~~ 34 University of Pennsylvania undergraduates to listen only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants**.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), $F(1, 18) = 1.01$, $p = .33$.

Approaches to a Solution

Recommendations (Wagenmakers et al., 2012)

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 632–638
©The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691612463078
http://pps.sagepub.com
SAGE

An Agenda for Purely Confirmatory Research

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit
University of Amsterdam, The Netherlands

Abstract

The veracity of substantive research claims hinges on the way experimental data are collected and analyzed. In this article, we discuss an uncomfortable fact that threatens the core of psychology's academic enterprise: almost without exception, psychologists do not commit themselves to a method of data analysis *before* they see the actual data. It then becomes tempting to fine tune the analysis to the data in order to obtain a desired result—a procedure that invalidates the interpretation of the common statistical tests. The extent of the fine tuning varies widely across experiments and experimenters but is almost impossible for reviewers and readers to gauge. To remedy the situation, we propose that researchers preregister their studies and indicate in advance the analyses they intend to conduct. Only these analyses deserve the label "confirmatory," and only for these analyses are the common statistical tests valid. Other analyses can be carried out but these should be labeled "exploratory." We illustrate our proposal with a confirmatory replication attempt of a study on extrasensory perception.

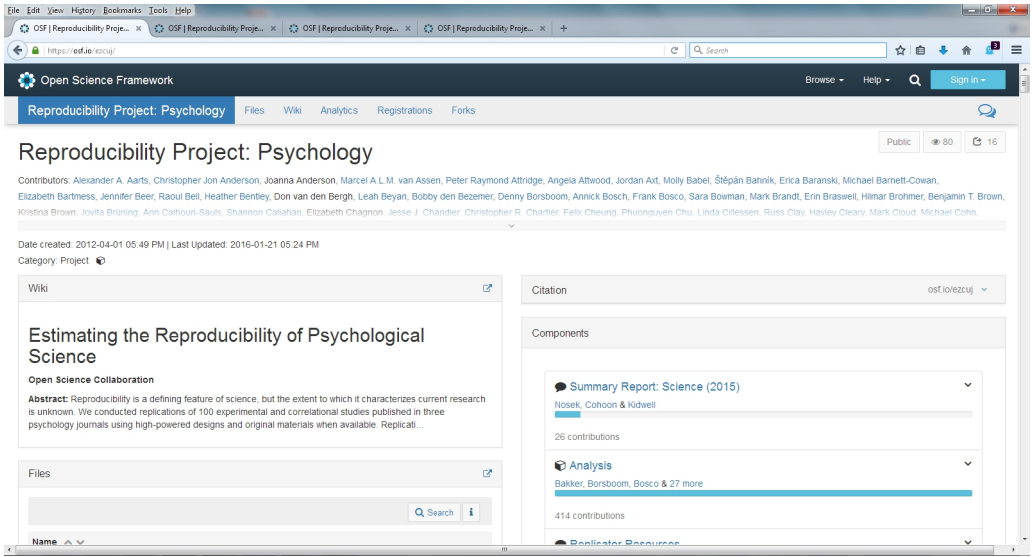
Approaches to a Solution

Recommendations by Michael Frank (`https://babieslearninglanguage.blogspot.com/2015/08/the-slower-harder-ways-to-increase.html`)

- Cumulative study sets with internal replication
- Everything open by default
- Larger N (based on realistic estimates of effect size, e. g., $\times 2.5$)
- Better visualizations of variability
- Manipulation checks, positive controls, and negative controls
- Predictions from (quantitative) theory
- A broader portfolio of publication outlets

Conclusions. A theme running throughout all these recommendations is that I generally believe in all of the principles advocated by folks who want prereg, publication of nulls, and Bayesian data analysis. But their suggestions are almost always posed as mechanical rules: rules that you can follow and know that you are doing science the right way. But these rules should be tempered by our judgment. Preregister if you think that previous work doesn't sufficiently constrain analysis. Publish nulls if they are theoretically important. Use Bayesian tools if they afford some analytic advantages relative to their complexity. But don't do these things just because someone said to. Do them because – in your best scientific judgment – they improve the reliability and validity of the argument you are making.

Reproducibility in Psychology



https://osf.io/ezcuj/

Reproducibility Project

Volunteer researchers selected a study to replicate from a quasi-random sample of studies published in three prominent journals from the 2008 publication year. Several strategies were followed to maximize the quality – and hence diagnosticity – of replication attempts. Each replication used sample sizes that achieved at least 80% power to detect an effect as large as the effect reported in the original article. Most studies achieved power levels of 90% or higher. Replication teams contacted original authors to acquire original materials and any missing methodological details required for a fair replication attempt. In addition, each team drafted a replication protocol outlining the procedure and data-analysis plan. The protocol followed a standard template and was sent to original authors for review. Any design discrepancies identified by the original authors were addressed or documented. Prior to data collection, the final replication protocol was registered and added to an online repository.

`https://osf.io/ezcuj/`

Reproducibility Project

Joint initiative

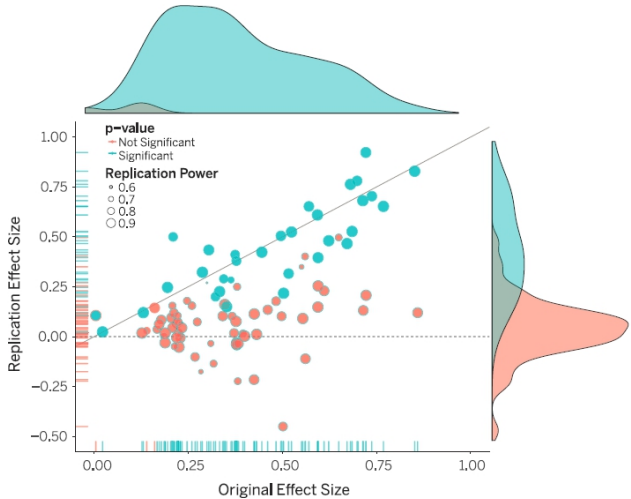
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349, aac4716. https://doi.org/10.1126/science.aac4716

conducted high-powered and preregistered replications of 100 experimental and correlational studies published in three psychology journals

Results

- The mean effect size of the replication effects was half the magnitude of the mean effect size of the original effects
- While 97% of the original studies had significant results, this holds for 36% of the replications
- 39% of effects were subjectively rated to have replicated the original result

Reproducibility Project



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.

Reproducibility Project

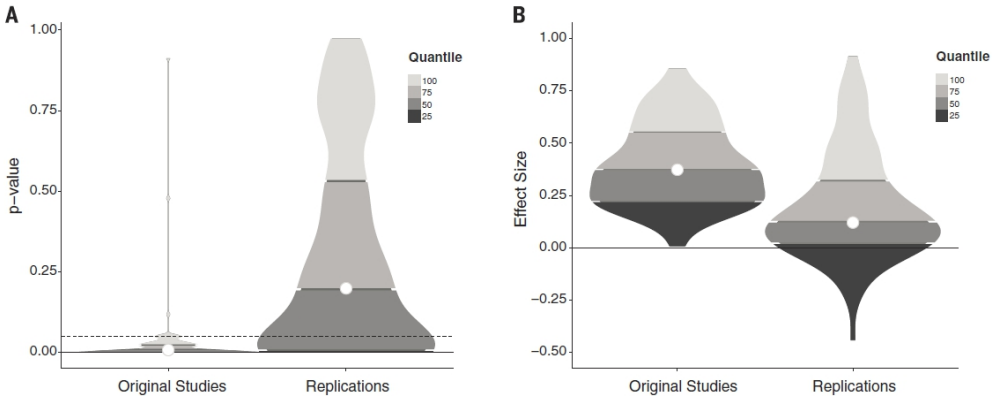


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Light at the End of the Tunnel . . .

The Brian Wansink case

- Former executive director of USDA's Center for Nutrition Policy and Promotion (CNPP) (2007-2009), and professor in the Applied Economics and Management Department at Cornell University
- As head of the *Cornell Food and Brand Lab* he conducted studies on eating behavior, was responsible for various food-related programs, and lead the US dietary guidelines
- In 2017 Wansink's research came under scrutiny
- After Wansink has written the following blog post, investigations of several related publications (referred to as "the pizza papers") were initiated

The Brian Wansink Case

A PhD student from a Turkish university called to interview to be a visiting scholar for 6 months. Her dissertation was on a topic that was only indirectly related to our Lab's mission, but she really wanted to come and we had the room, so I said "Yes."

When she arrived, I gave her a data set of a self-funded, failed study which had null results (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people 1/2 as much as others). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that held up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up, and every day for a month I told her how to rewrite it and she did. This happened with a second paper, and then a third paper (which was one that was based on her own discovery while digging through the data). . . .

The Brian Wansink Case

. . . At about this same time, I had a second data set that I thought was really cool that I had offered up to one of my paid post-docs (again, the woman from Turkey was an unpaid visitor). In the same way this same post-doc had originally declined to analyze the buffet data because they weren't sure where it would be published, they also declined this second data set. They said it would have been a "side project" for them they didn't have the personal time to do it. Boundaries. I get it.

Six months after arriving, the Turkish woman had one paper accepted, two papers with revision requests, and two others that were submitted (and were eventually accepted – see below). In comparison, the post-doc left after a year (and also left academia) with 1/4 as much published (per month) as the Turkish woman. I think the person was also resentful of the Turkish woman.

Balance and time management has its place, but sometimes it's best to "Make hay while the sun shines." . . .

The Brian Wansink Case

. . . About the third time a mentor hears a person say "No" to a research opportunity, a productive mentor will almost instantly give it to a second researcher – along with the next opportunity. This second researcher might be less experienced, less well trained, from a lessor school, or from a lessor background, but at least they don't waste time by saying "No" or "I'll think about it." They unhesitatingly say "Yes" – even if they are not exactly sure how they'll do it.

Facebook, Twitter, Game of Thrones, Starbucks, spinning class . . . time management is tough when there's so many other shiny alternatives that are more inviting than writing the background section or doing the analyses for a paper.

Yet most of us will never remember what we read or posted on Twitter or Facebook yesterday. In the meantime, this Turkish woman's resume will always have the five papers below.

`https://web.archive.org/web/20170312041524/http:`
`/www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no`

The Brian Wansink Case

- Closer inspection of Brian Wansink's publications revealed various problems
  - Conclusions not supported by the data presented
  - Data and figures duplicated across papers
  - Questionable data (including impossible values)
  - Incorrect and inappropriate statistical analyses
  - P-hacking
- As of 2020 Wansink has had 18 of his research papers retracted (one paper retracted twice)
- In September 2018 Cornell determined that Wansink had committed scientific misconduct and removed him from research and teaching activities
- Wansink resigned effective June 30, 2019

The Brian Wansink Case

Resources

- Overview
  https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking
- Analysis of Wansink's publications by Tim van der Zee, Jordan Anaya and Nicholas J.L. Brown
  https://bmcnutr.biomedcentral.com/articles/10.1186/s40795-017-0167-x
- Statement of Cornell University
  https://statements.cornell.edu/2018/20180920-statement-provost-michael-kotlikoff.cfm
- Youtube video "Sloppy Food Science"
  https://www.youtube.com/watch?v=u1I6p851rLI