p-hacking:
What it is, how to prevent it

This presentation is licensed under a CC-BY 4.0 license. You may copy, distribute, and use the slides in your own work, as long as you give attribution to the original author at each slide that you use.
Much of the scientific literature, perhaps half, may simply be untrue.

Part of the problem is that no one is incentivised to be right.
Researchers are not rewarded for being right, but rather for publishing a lot.

Nelson, Simmons, & Simonsohn (2012); Nosek, Spies, Motyl (2012); Munafo (2016)
How can we publish a lot?

Psychology/Psychiatry

92%!

\( p \)-hack your way to scientific glory!
**p-hacking** (*n.*) Tune your data analysis in a way that you achieve a significant *p*-value in situations where it would have been non-significant.

**Questionable research practices (QRPs) ** (*n.*) Practices of data collection and data analysis that are not outright fraud, but also not really kosher.
Tool 1: Outcome switching

Tracking switched outcomes in clinical trials

Here’s what we found.

<table>
<thead>
<tr>
<th></th>
<th>67</th>
<th>9</th>
<th>300</th>
<th>357</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials checked</td>
<td>Trials were perfect</td>
<td>Outcomes not reported</td>
<td>New outcomes silently added</td>
<td></td>
</tr>
</tbody>
</table>

On average, each trial reported just 62.1% of its specified outcomes. And on average, each trial silently added 5.3 new outcomes.

Tool 1: Outcome switching

• 2 outcome variables:
  false positive rate 5% → 9.5%

• 5 outcome variables with one-sided testing:
  false positive rate 5% → 41%

For Vohs et al. (2006), “the authors conducted two additional money priming studies that showed no effects, the details of which were shared with us.” and “reported nine dependent measures that were statistically affected by the manipulation in the predicted direction (one in each experiment) but did not report 19 additional measures that were statistically unchanged”.
Tool 2: Many conditions, report only those that worked

Best-practice example: Transform a boring dissertation into a groundbreaking publication

https://twitter.com/JoeHilgard/status/699693258386051072

Here's another spicy one: Thesis reports four conditions, 415 subjects. Manuscript reports three conditions, 140 subjects.
### Tool 3: Optional stopping

**Table 2**

The probability of being absorbed at or before the nth observation in sampling from a normal distribution with known variance, with repeated tests at a nominal two-sided significance level $2\alpha$ (i.e. standardized normal deviate $k$)

<table>
<thead>
<tr>
<th>$2\alpha$ k</th>
<th>0·10 1·645</th>
<th>0·05 1·960</th>
<th>0·02 2·326</th>
<th>0·01 2·576</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Q</td>
<td>S</td>
<td>Q</td>
<td>S</td>
</tr>
<tr>
<td>1</td>
<td>0·10000</td>
<td>0·0970</td>
<td>0·05000</td>
<td>0·0545</td>
</tr>
<tr>
<td>2</td>
<td>0·16015</td>
<td>0·1650</td>
<td>0·08312</td>
<td>0·0885</td>
</tr>
<tr>
<td>3</td>
<td>0·20207</td>
<td>0·1980</td>
<td>0·10726</td>
<td>0·1115</td>
</tr>
<tr>
<td>4</td>
<td>0·23399</td>
<td>0·2295</td>
<td>0·12617</td>
<td>0·1260</td>
</tr>
<tr>
<td>5</td>
<td>0·25963</td>
<td>0·2590</td>
<td>0·14169</td>
<td>0·1420</td>
</tr>
<tr>
<td>160</td>
<td>0·63315</td>
<td></td>
<td>0·40829</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>0·64301</td>
<td></td>
<td>0·41677</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0·65165</td>
<td></td>
<td>0·42429</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0·670</td>
<td></td>
<td>0·440</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0·720</td>
<td></td>
<td>0·487</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>0·746</td>
<td></td>
<td>0·513</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>0·763</td>
<td></td>
<td>0·529</td>
<td></td>
</tr>
</tbody>
</table>

*With long enough sampling and optional stopping, it is guaranteed to get a significant result!*}

Tool 4: Analytical flexibility

Quantification Strategies

While the CRT ostensibly measures how much unpleasant, or even harmful, noise a participant is willing to administer to a nonexistent confederate, that amount of noise can be extracted as a measure in myriad different ways using various combinations of volume and duration over one or more trials. There are currently 117 publications that report using the CRT stored in the database, and they reported 142 different quantification strategies in total.

The number of quantification strategies appears to be growing proportionally to the number of publications that use the CRT, and at no point in time since its first use were there more publications than different quantification strategies. That is, although the total number of unique quantification strategies has grown substantially in 30 years, new ones keep being added.

Publications and Quantification Strategies Accumulated Growth
The table below lists all (co-)authors from all publications that report at least one CRTT quantification strategy.
Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling IN PSYCHOLOGY

Leslie K. John¹, George Loewenstein², and Drazen Prelec³

¹Marketing Unit, Harvard Business School; ²Department of Social & Decision Sciences, Carnegie Mellon University; and ³Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology
Pro-Tip: Build the p-hacking into the software!
How bad is it?

http://shinyapps.org/apps/p-hacker/
I believe that you should collectively do something about this mess.

I see a train wreck looming.

Daniel Kahneman, Nobel prize 2002

http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf
Which part of published research can be independently replicated?

*The data on economics is about *reproducibility*; i.e. the attempt to get the same results if you apply the original data analysis on the original data set.*

---

IS THERE A REPRODUCIBILITY CRISIS?

7% Don’t know
52% Yes, a significant crisis
38% Yes, a slight crisis
3% No, there is no crisis

1,576 researchers surveyed

90%: Yes

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
Retractions: +1000% in 10 years

“In the past decade, the number of retraction notices has shot up 10-fold.”

Scientific misconduct:

+ **1200%** in 4 years

U.S. Office of Research Integrity:
Confirmed cases of scientific misconduct

2009-2011: 3 cases
2012-2015: 36 cases

https://ori.hhs.gov/case_summary
„Innovative, unprecedented, transformative!“
+880% from 1974-2014

**Thesis:**

Our current incentives foster questionable research practices, which decrease the truth value of our shared knowledge.

What is good for the individual careers of researchers leads to a collective fiasko.

Researchers who do it right (i.e., high power, no QRPs, transparency) have a clear competitive disadvantage.

---

**Anti-Thesis:**

Society pays for us that we generate valid and robust knowledge.

Our incentives should be chosen in a way that they foster good science.

Researchers who do it right should be supported and promoted.
How to prevent $p$-hacking

1. Reduce analytical degrees of freedom
2. Embrace analytical degrees of freedom
3. Change the incentives
How to prevent $p$-hacking

1. **Reduce analytical degrees of freedom**
2. Embrace analytical degrees of freedom
3. Change the incentives
Registered Reports

https://cos.io/rr/
Registered Reports

https://cos.io/rr/

82 journals

offer Registered Reports

(by December 2017)

AIMS Neuroscience
Attention, Perception, and Psychophysics
Cognition and Emotion
Comprehensive Results in Social Psychology
Cortex
Drug and Alcohol Dependence
Experimental Psychology

Perspectives on Psychological Science
Royal Society Open Science
Social Psychology
Working, Aging and Retirement
UPR: unreviewed pre-reg

RPR/RR: reviewed pre-reg/Registered Reports

Private
UPR
unreviewed pre-reg

RPR/RR
reviewed pre-reg/
Registered Reports

private

non-timestamped
(internal only)

any physical or
digital file drawer of
your choice

osf.io/view/psyarxiv

Slides by Anne Scheel
### Where *can I preregister*?

<table>
<thead>
<tr>
<th>RPR</th>
<th>Registered Reports</th>
<th>✓</th>
<th>&quot;gold standard&quot;</th>
<th>not offered by every journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPR</td>
<td>internal only</td>
<td>–</td>
<td>better planning of studies</td>
<td>not externally recognised</td>
</tr>
<tr>
<td></td>
<td>private online repository</td>
<td>– / ✓ (depends)</td>
<td>no scooping</td>
<td>not hack-safe (misconduct), doesn't prevent file-drawer effect</td>
</tr>
<tr>
<td></td>
<td>public online repository supporting public PR, e.g. OSF</td>
<td>✓</td>
<td>find collaborators, get feedback, file-drawer safe</td>
<td>Scooping possible? (But: time-stamp proofs date of creation; scooping prevented by embargo)</td>
</tr>
</tbody>
</table>

→ all UPRs can be done with and without a template (exception: AsPredicted is template-only)
no prereg: 57% success rate!

prereg: 8% success rate...

How to prevent $p$-hacking

1. **Reduce analytical degrees of freedom**
2. Embrace analytical degrees of freedom
3. Change the incentives

Prevent $p$-hacking by reducing the data analysis to a single pipeline.
How to prevent \( p \)-hacking

1. Reduce analytical degrees of freedom

2. Embrace analytical degrees of freedom

3. Change the incentives
Many names for the same idea …

• Sensitivity/ robustness analysis
• Multiverse analysis (Steegen et al., 2016)
• Specification curve (Simonsohn et al., 2015)
• Vibration of effects (Patel et al., 2015)
• Ensemble approach (e.g. climatology)
  ➔ use a set of models with the same input data to produce a range of outcomes

cf. Boulesteix, Elsas, & Schönbrodt (in prep)
• A “multiverse analysis” (Steegen, Tuerlinchx, Gelman, & Vanpaemel, 2016): Report results for all plausible analytical decisions

• Check robustness of results: Do several analytical paths lead to comparable conclusions?

• Based on open data by Carney et al. (2010)
Table 1. Multiverse Analysis for the Effect of Power Posing on Testosterone.

<table>
<thead>
<tr>
<th>Gender Effect</th>
<th>Control Variables</th>
<th>Outlier Identification: Entire Sample (N = 39)</th>
<th>Outlier Identification: Test. Conditioned on Gender (N = 41)</th>
<th>Outlier Identification: Multivariate or No Exclusion (N = 42)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>Gender</td>
<td>.047 (p = .19)</td>
<td>.019 (p = .39)</td>
<td>.036 (p = .23)</td>
</tr>
<tr>
<td>Combined</td>
<td>Gender and T1 test.</td>
<td>.029 (p = .31)</td>
<td>.042 (p = .21)</td>
<td>.055 (p = .15)</td>
</tr>
<tr>
<td>Combined</td>
<td>Gender, T1 test., and T1 cort.</td>
<td>.037 (p = .26)</td>
<td>.017 (p = .43)</td>
<td>.018 (p = .42)</td>
</tr>
<tr>
<td>Combined</td>
<td>T1 cort. and T2 cort.</td>
<td>.089 (p = .07)</td>
<td>.038 (p = .23)</td>
<td>.037 (p = .24)</td>
</tr>
<tr>
<td>Combined</td>
<td>Gender, T1 test., T1 cort., and T2 cort.</td>
<td><strong>.123 (p = .04)</strong></td>
<td>.099 (p = .06)</td>
<td>.102 (p = .051)</td>
</tr>
<tr>
<td>Men only</td>
<td>No controls</td>
<td>.192 (p = .13)</td>
<td>.047 (p = .44)</td>
<td>.096 (p = .24)</td>
</tr>
<tr>
<td>Men only</td>
<td>T1 test.</td>
<td>.000 (p = .96)</td>
<td>.073 (p = .35)</td>
<td>.101 (p = .25)</td>
</tr>
<tr>
<td>Men only</td>
<td>T1 cort.</td>
<td>.184 (p = .17)</td>
<td>.121 (p = .22)</td>
<td>.063 (p = .37)</td>
</tr>
<tr>
<td>Men only</td>
<td>T1 test. and T1 cort.</td>
<td>.026 (p = .64)</td>
<td>.104 (p = .28)</td>
<td>.083 (p = .32)</td>
</tr>
<tr>
<td>Men only</td>
<td>T1 cort. and T2 cort.</td>
<td>.162 (p = .22)</td>
<td>.141 (p = .21)</td>
<td>.057 (p = .41)</td>
</tr>
<tr>
<td>Men only</td>
<td>T1 test., T1 cort., and T2 cort.</td>
<td>.026 (p = .66)</td>
<td>.125 (p = .26)</td>
<td>.086 (p = .33)</td>
</tr>
<tr>
<td>Women only</td>
<td>No controls</td>
<td>.005 (p = .73)</td>
<td>.005 (p = .73)</td>
<td>.005 (p = .73)</td>
</tr>
<tr>
<td>Women only</td>
<td>T1 test.</td>
<td>.019 (p = .51)</td>
<td>.019 (p = .51)</td>
<td>.019 (p = .51)</td>
</tr>
<tr>
<td>Women only</td>
<td>T1 cort.</td>
<td>.005 (p = .75)</td>
<td>.005 (p = .75)</td>
<td>.005 (p = .75)</td>
</tr>
<tr>
<td>Women only</td>
<td>T1 test. and T1 cort.</td>
<td>.023 (p = .48)</td>
<td>.023 (p = .48)</td>
<td>.023 (p = .48)</td>
</tr>
<tr>
<td>Women only</td>
<td>T1 cort. and T2 cort.</td>
<td>.077 (p = .19)</td>
<td>.077 (p = .19)</td>
<td>.077 (p = .19)</td>
</tr>
<tr>
<td>Women only</td>
<td>T1 test., T1 cort., and T2 cort.</td>
<td>.167 (p = .053)</td>
<td>.167 (p = .053)</td>
<td>.167 (p = .053)</td>
</tr>
</tbody>
</table>

Note. Entries are partial η² values and (in parentheses) the associated p value. The entry in boldface is the effect for the analyses originally reported in the Carney, Cuddy, and Yap (2010) paper. Blank entries mean that the analyses would not be recommended for reasons described in the text. The number of women was constant across the three outlier strategies. DV = dependent variable; Test. = testosterone; cort. = cortisol; T1 = premanipulation; T2 = postmanipulation.

Of 54 plausible analyses exactly **one** was significant.
Guess which has been reported in the original paper?
5. Initially, the primary DV of interest was risk-taking. We ran subjects in chunks and checked the effect along the way. It was something like 25 subjects run, then 10, then 7, then 5. Back then this did not seem like p-hacking. It seemed like saving money (assuming your effect size was big enough and p-value was the only issue).

6. Some subjects were excluded on bases such as “didn’t follow directions.” The total number of exclusions was 5. The final sample size was $N = 42$.

7. The cortisol and testosterone data (in saliva at that point) were sent to Salimetrics (which was in State College, PA at that time). The hormone results came back and data were analyzed.

8. For the risk-taking DV: One p-value for a Pearson chi square was .052 and for the Likelihood ratio it was .05. The smaller of the two was reported despite the Pearson being the more ubiquitously used test of significance for a

10. The self-report DV was p-hacked in that many different power questions were asked and those chosen were the ones that “worked.”
Probing Birth-Order Effects on Narrow Traits Using Specification Curve Analysis

Julia M. Rohrer¹,², Boris Egloff⁵, Stefan C. Schmukle²

---

Prerequisite: Open Data

“dass die Daten unmittelbar nach Abschluss der Forschungen oder nach wenigen Monaten der Öffentlichkeit frei zur Verfügung gestellt werden.“

„Das Engagement […] von Wissenschaftlern und Wissenschaftlerinnen um die Verfügbarmachung von Forschungsdaten sollten bei der Würdigung von wissenschaftlichen […] Leistungen zukünftig stärker berücksichtigt werden“

---

Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien

Im Auftrag des DGPs Vorstands (17.09.2016)

Felix Schönbrodt, Mario Gollwitzer und Andrea Abele-Brehm


Journals with mandatory open data (or justification why not)

• Advances in Methods and Practices in Psychological Science
  (http://www.psychologicalscience.org/publications/ampps/ampps-submission-guidelines#DISC)

• Collabra: Psychology
  (https://www.collabra.org/about/research-integrity/)

• Experimental Psychology

• Journal of Research in Personality
  (http://www.sciencedirect.com/science/article/pii/S0092656617300211)

• Judgment and Decision Making
  (http://journal.sjdm.org/)

• Journal of Cognition
  (https://www.journalofcognition.org/about/editorialpolicies/)

• PLOS ONE
  (http://blogs.plos.org/everyone/2017/05/08/making-progress-toward-open-data/)

• Royal Society Open Science
  (http://rsos.royalsocietypublishing.org/author-information#Open_data)

• Science
  (http://www.sciencemag.org/authors/science-editorial-policies)
How to prevent $p$-hacking

1. Reduce analytical degrees of freedom
2. **Embrace analytical degrees of freedom**
3. Change the incentives

Show robustness against $p$-hacking by computing all (sensible) analytical pipelines
How to prevent \( p \)-hacking

1. Reduce analytical degrees of freedom
2. Embrace analytical degrees of freedom
3. **Change the incentives**
Proximate incentives

• Why do we $p$-hack? To get a publishable result.

• Hypothesis: Publication bias as the driving force behind $p$-hacking.
  
  • Carter, Schönbrodt, Gervais, & Hilgard (2017): $p$-hacking (without publication bias) has minor impact; but quite problematic in combination with publication bias (see https://psyarxiv.com/9h3nu/)

• But also see Simonsohn (2016):
  
  • “$P$-hacking is easy to stop. File-drawering nearly impossible. Fortunately, while $p$-hacking is a real problem, file-drawering is not.“

• http://datacolada.org/55
The Department of Psychology at the Faculty of Human Sciences of the University of Cologne (UoC) seeks to appoint a Full Professor (W3) of Social Psychology to be filled as soon as possible.

The successful candidate is expected to have a record of excellence in social cognition, and/or related areas such as cognitive psychology or motivation science. The candidate is also expected to strongly contribute to the UoC's Center for Social and Economic Behavior and the Social Cognition Center Cologne of the Department of Psychology. Both structures are part of UoC's Key Profile Area II, "Behavioral Economic Engineering and Social Cognition".

The ideal candidate's track record should show an excellent fit with these interrelated structures and a strong interest to bridge the fields of social cognition and behavioral economics.

The Department of Psychology aims for transparent and reproducible research (including Open Data, Open Materials, and Preregistrations). Applicants are asked to illustrate how they have pursued these goals in the past and/or how they plan to do so in the future.

We strongly encourage international applicants. Salaries and working conditions at the UoC – one of the German Universities of Excellence – meet international standards. Candidates are expected to be willing to learn the German language. The Faculties offer Bachelor, Master, and doctoral degrees. Courses are taught either in English or German.

Applicants will be hired in concordance with § 36 of the University Law of the State of North Rhine Westphalia.

The UoC supports diversity, the multiplicity of perspectives, and equal opportunities. The University of Cologne particularly encourages applications from disabled persons. Disabled persons are given preference in case of equal qualification. Women are strongly encouraged to apply. Preferential treatment is given to women if their professional qualifications and abilities are equivalent to those of other applicants.

Applications with the usual documents (including vita, research statement, 5 most important publications, full list of publications and teaching experience, and diplomas) should be submitted via the University's Academic Job Portal (https://berufungen.uni-koeln.de) until March 30th, 2017.

https://osf.io/dbkva/
• 17 members of 10 disciplines:
  Psychology, sociology, computer science, statistics, geography, medicine, veterinary medicine, economics, ...

• 3 entire faculties as members:
  Faculty of Medicine, Faculty of Veterinary Medicine, Faculty of Psychology and Educational Science

• Mission Statement:
  • Education (from PhD student to professor)
  • Meta-science research
  • **Change the incentive structure**

• http://www.osc.lmu.de
Thanks for your attention!