Assignment

Misinterpretation of *p*-values

Ian Hussey

Digitalisation of Psychology

Misinterpretations of *p* values

Non-significant *p* values

Comparing *p* values

Misinterpretations of *p* values

Non-significant *p* values

Absence of evidence does not equal evidence of absence

Abwesenheit von Evidenz für einen Effekt ist nicht gleichbedeutend mit Evidenz für die Abwesenheit eines Effekts

Comparing *p* values

The difference between “statistically significant” and “statistically non-significant” is not itself significant

Der Unterschied zwischen „statistisch signifikant“ und „statistisch nicht signifikant“ ist selbst nicht statistisch signifikant.

Misinterpretations of *p* values
Results

Non-significant *p* values

47% of articles reporting this test have 1+ misinterpretations

evidence_for_null_error	n_tests_evidence_for_null_error	percent_error
FALSE	31	53.4
TRUE	27	46.6

Comparing *p* values

25% of articles reporting this test have 1+ misinterpretations

evidence_of_moderation_error	n_tests_evidence_of_moderation_error	percent_error
FALSE	37	75.5
TRUE	12	24.5

Misinterpretations of *p* values
Results

Non-significant *p* values

QUIZ

More misinterpretation of *p*-values

Check your understanding

Are these correct interpretations?

1. *p*-value is the probability that the test hypothesis is true
2. *p*-value is the probability that chance alone produced the effect
3. *p*-value is the the probability the effect will replicate (1 - *p*)
4. Statistical significance = a scientifically or substantively important discovery
5. $p < .05$ means the alternative hypothesis is true
6. $p >= .05$ means the null hypothesis is true

Lakens - Improving Your Statistical Inferences

Überprüfen Sie Ihr Verständnis

Sind dies die richtigen Interpretationen?

1. Ein *p*-value ist die Wahrscheinlichkeit, dass die Testhypothese wahr ist.
2. Ein *p*-value ist die Wahrscheinlichkeit, dass der Effekt allein durch Zufall entstanden ist.
3. Ein *p*-value ist die Wahrscheinlichkeit, dass der Effekt in einer zukünftigen Studie repliziert werden kann (1 - *p*).
4. Ein statistisch signifikantes Ergebnis impliziert eine wissenschaftlich oder inhaltlich wichtige Entdeckung.
5. Wenn $p < .05$ ist, bedeutet dies, dass die Alternativhypothese wahr ist.
6. Wenn $p >= .05$ ist, bedeutet dies, dass die Nullhypothese wahr ist.

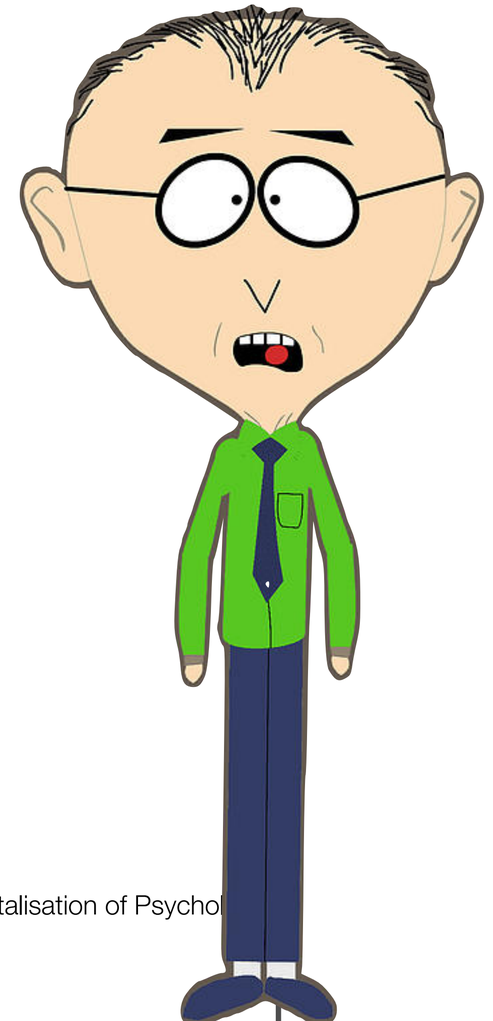
Lakens - Improving Your Statistical Inferences

Understanding the long run of *p* values

PPV of p-values

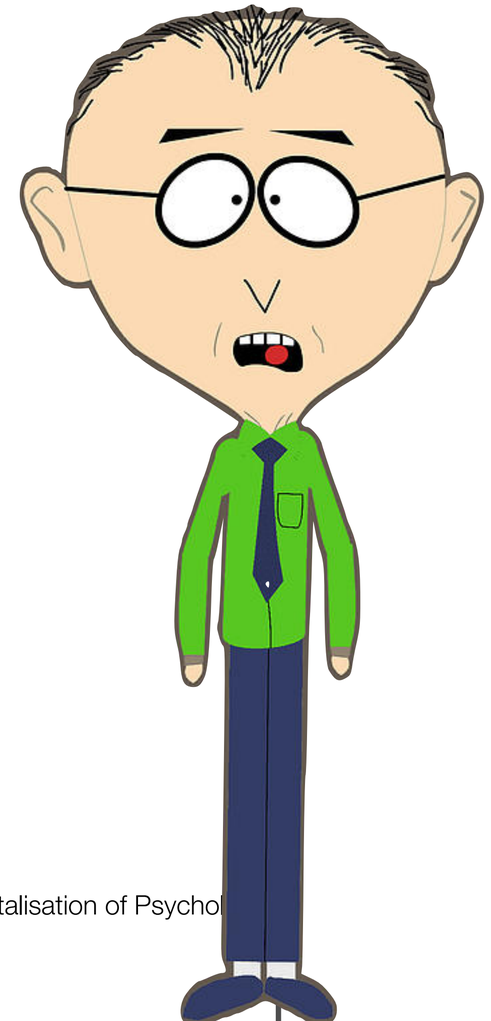
p-hacking & credibility

p-hacking is bad



p-hacking is bad

Selective reporting DV
Selective reporting IV
Optional stopping outlier exclusion
Controlling covariates
Scale redefinition
Variable transformation
Discretizing variables
Alternative hypothesis tests
Favorable imputation
Subgroup analyses
Incorrect rounding



p-hacking is bad
Which forms are worst?

Selective reporting DV

Selective reporting IV

Optional stopping outlier exclusion

Controlling covariates

Scale redefinition

Variable transformation

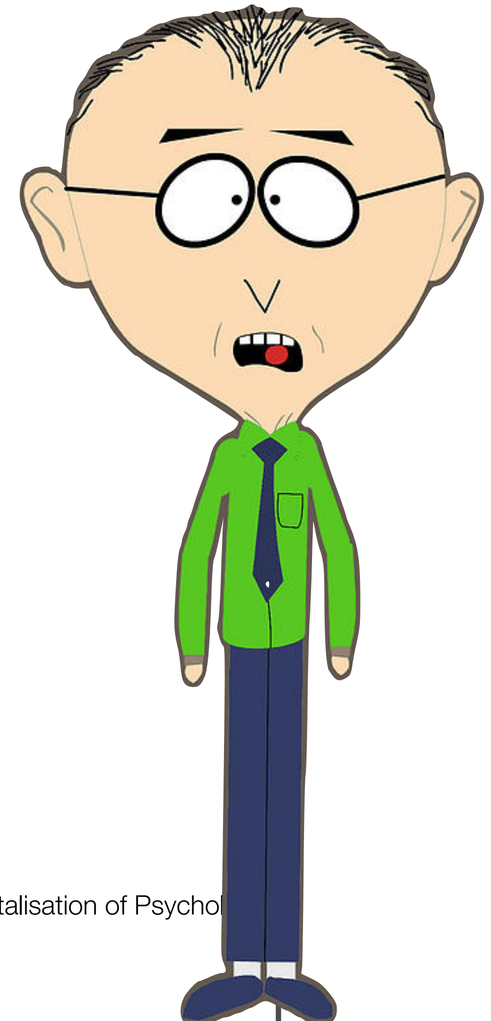
Discretizing variables

Alternative hypothesis tests

Favorable imputation

Subgroup analyses

Incorrect rounding



p-hacking is bad
Which forms are worst?

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Stefan AM, Schönbrodt FD. 2023 Big little lies: a compendium and simulation of *p*-hacking strategies. *R. Soc. Open Sci.* **10**: 220346.
https://doi.org/10.1098/rsos.220346

Received: 17 March 2022
Accepted: 11 January 2023

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:

Big little lies: a compendium and simulation of *p*-hacking strategies

Angelika M. Stefan[1,2] and Felix D. Schönbrodt[3]

[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
[2]Department of Psychology, Universität der Bundeswehr München, München, Germany
[3]Department of Psychology, Ludwig-Maximilians-Universität München, Munchen, Germany

AMS, 0000-0003-3382-4746

In many research fields, the widespread use of questionable research practices has jeopardized the credibility of scientific results. One of the most prominent questionable research practices is *p*-hacking. Typically, *p*-hacking is defined as a compound of strategies targeted at rendering non-significant hypothesis testing results significant. However, a comprehensive overview of these *p*-hacking strategies is missing, and current meta-scientific research often ignores the heterogeneity of strategies. Here, we compile a list of 12 *p*-hacking strategies based on an extensive literature review, identify factors that control their level of severity, and demonstrate their impact on false-positive rates using simulation studies. We also use our simulation results to evaluate several approaches that have been proposed to mitigate the influence of questionable research practices. Our results show that investigating *p*-hacking at the

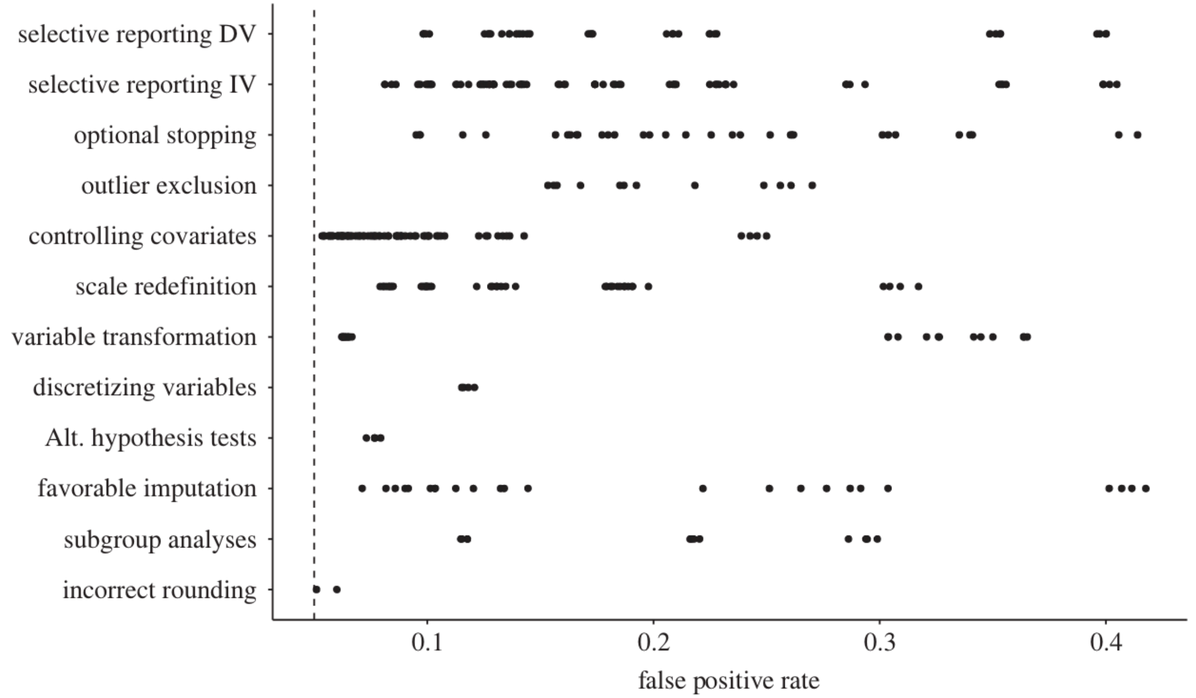


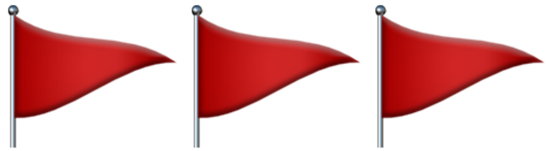
Figure 12. Overview of *p*-hacking severity in terms of false-positive rates for different all *p*-hacking strategies discussed in this paper.

p-hacking is bad

Not useful for assessing individual articles!

Why?

Statistical red flags



p-hacking is bad

Not useful for assessing individual articles!

Why?
You don’t have access to the information you need:
- Everything that was done
- Everything that wasn’t done
- When

Be aware of *p*-hacking when thinking about credibility

But also there is rarely a test for whether *p*-hacking occurred

What *can* you look for?
A checklist of statistical red flags

- Reliance on *p* values between between .02 and .05, especially if one-tailed
- Reliance on *p* values only with no effect sizes (standardized or unstandardized).
- Subgroup analyses
- Use of mediation analysis (or variants, eg mediated moderation, moderated mediation, etc: use of the PROCESS macro generally)
- Use of MANOVA
- 3+ way interactions in ANOVA/ANCOVA/MANOVA

What *can* you look for?
A checklist of statistical red flags

- Use of Stepwise Regression
- Covariate selection on the basis of *p* values or effect sizes rather than causal justification
- Post hoc power analysis
- A priori power analysis but for a different test than the one conducted (e.g., power analysis for ANOVA but ran mixed models)
- Conditioning on a post-treatment variable in a randomized experiment, eg excluding on the basis of a manipulation check

Wonach können Sie suchen?
Eine Checkliste mit statistischen Warnsignalen

- Verlassen auf *p*-values zwischen 0.02 und 0.05, insbesondere bei einseitiger
- Verlassen auf *p*-values nur ohne Effektgrößen (standardisiert oder nicht standardisiert).
- Untergruppenanalysen
- Verwendung von Mediationsanalysen (oder Varianten, z. B. mediated moderation, moderated mediation usw.: Verwendung des Makros PROCESS im Allgemeinen)
- Verwendung von MANOVA
- 3+ Interaktionen in ANOVA/ANCOVA/MANOVA

Wonach können Sie suchen?
Eine Checkliste mit statistischen Warnsignalen

- Verwendung der schrittweisen Regression
- Kovariablenauswahl auf der Grundlage von p-Werten oder Effektgrößen anstelle einer kausalen Begründung
- Post-hoc-Poweranalyse
- A-priori-Poweranalyse, aber für einen anderen Test als den durchgeführten (z. B. Poweranalyse für ANOVA, aber gemischte Modelle)
- Konditionierung auf eine Nachbehandlungsvariable in einem randomisierten Experiment, z. B. Ausschluss auf der Grundlage einer Manipulationsprüfung

Readings

- Brown & Heathers (2017) The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology
  - Properly read
    - Understand how GRIM works
    - Understand their findings from applying GRIM
- Heathers (2024) Approximately 1 in 7 Scientific Papers Are Fake
  - Recall:
  - Brown & Heathers (2017) did not report a key result: “twelve (12) manuscripts contained both multiple inconsistencies, and the authors refused and/or ignored a request for data, within which three (3) manuscripts contained what we considered definite hallmarks of systematic manipulation. These figures were both sufficiently speculative and controversial at the time of publication to lead us to redact them from the manuscript.”

Assignment

Use the collaborative Google Sheet to code statistical red flags in your articles

- The articles assigned to you for StatCheck
- Full instructions will be on Ilias

Questions?