UNIVERSITY OF MANNHEIM

Detecting Evidential Value and *P*-Hacking With the *P*-curve tool: A Word of Caution

Edgar Erdfelder and Daniel W. Heck

STATISTICAL MODELING in PSYCHOLOGY

FREIBURG HEIDELBERG LANDAU MANNHEIM TÜBINGEN

The *P*-Curve Tool

Journal of Experimental Psychology: General
2014, Vol. 143, No. 2, 534–547

0096-3445/14/$12.00 DOI: 10.1037/a0033242

P-Curve: A Key to the File-Drawer

Uri Simonsohn
University of Pennsylvania

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

Because scientists tend to report only studies (publication bias) or analyses (*p*-hacking) that "work," readers must ask, "Are these effects true, or do they merely reflect selective reporting?" We introduce *p*-curve as a way to answer this question. *P*-curve is the distribution of statistically significant *p* values for a set of studies ($ps < .05$). Because only true effects are expected to generate right-skewed *p*-curves—containing more low (.01s) than high (.04s) significant *p* values—only right-skewed *p*-curves are diagnostic of evidential value. By telling us whether we can rule out selective reporting as the sole explanation for a set of findings, *p*-curve offers a solution to the age-old inferential problems caused by file-drawers of failed studies and analyses.

The problem

- The false-positive rate in published psychological research is larger than α = .05.
- It may even exceed .60 in prototypical APA journals (cf. Open Science Collaboration, 2015).
- Obviously, a selection process is at work that favors significant outcomes.
- Possible Reason 1:
  - Reviewers and journal editors prefer significant results.
- Possible Reason 2:
  - Authors analyze their data in a way that inflates the α error risk.

p-hacking

- Definition of *p*-hacking (Simmons, Nelson & Simonsohn, 2011):
  - Analyze your data in multiple ways and stop upon the first significant result consistent with your prediction.
- Examples of *p*-hacking:
  - *Data peeking*: Continue sampling until significant.
  - *Multiple testing*: Try different dependent variables.
  - *Data trimming*: Eliminate “outliers” or transform variables until a significant outcome is observed.
  - *Statistical refinements*: E.g., try different covariates
  - *HARKing*: **H**ypothesizing **A**fter the **R**esults are **K**nown.

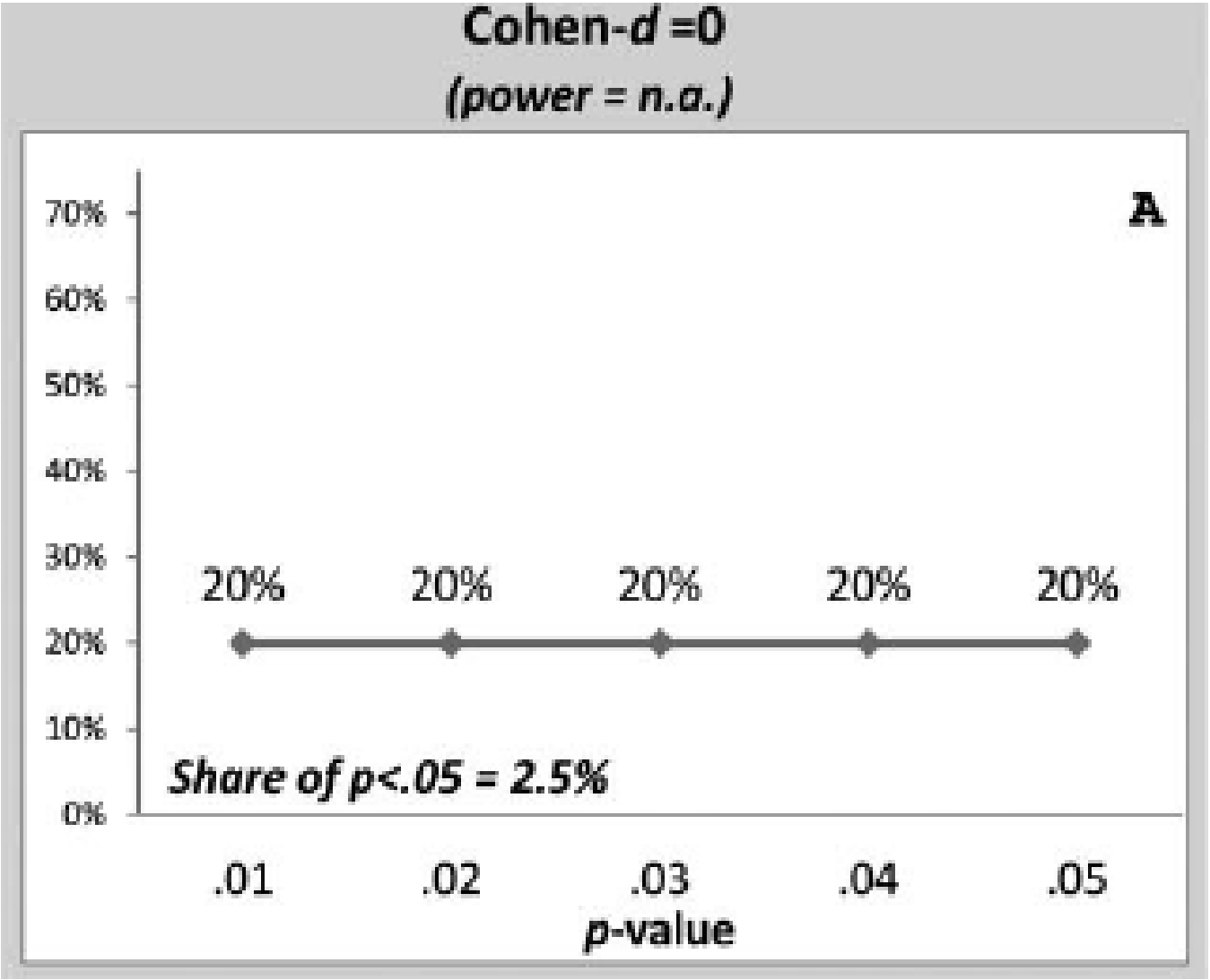
How to discriminate between true effects and false positive results?

- Goal:
  - Find a tool that discriminates between sets of significant results that are
    - (a) true positives (real effects)
    - (b) false positives (α errors or *p*-hacked null effects)
- Method (Simonsohn et al., 2014):
  - *P*-curve
  - Histogram of significant *p*-values from independent studies (hence, $p < .05$)
  - Basis: studies that meet a number of predefined inclusion criteria.

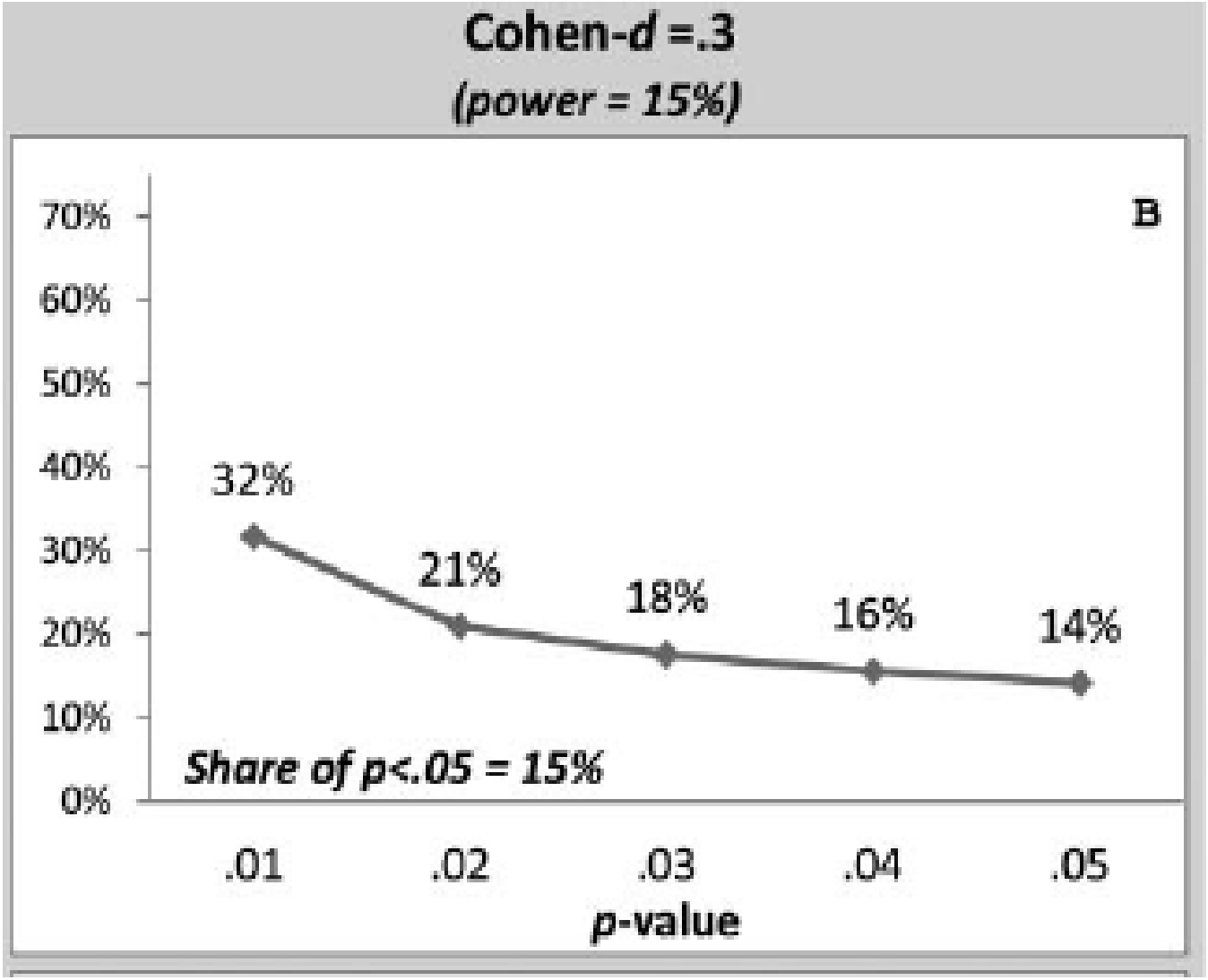
The role of evidential value

- No evidential value:
  - If H_0 holds in each of the tests included in the *p*-curve
  - and the underlying statistical model is valid
  - then *p*-values follow a uniform distribution
- Positive evidential value:
  - If H_1 holds in each of the tests included in the *p*-curve
  - and the underlying statistical model is valid
  - then p-values follow a right-skewed distribution.
  - The degree of skewness is a function of power under H_1.
  - More precisely, skewness depends on the noncentrality parameter under H_1 (in general: *N* x effect size)

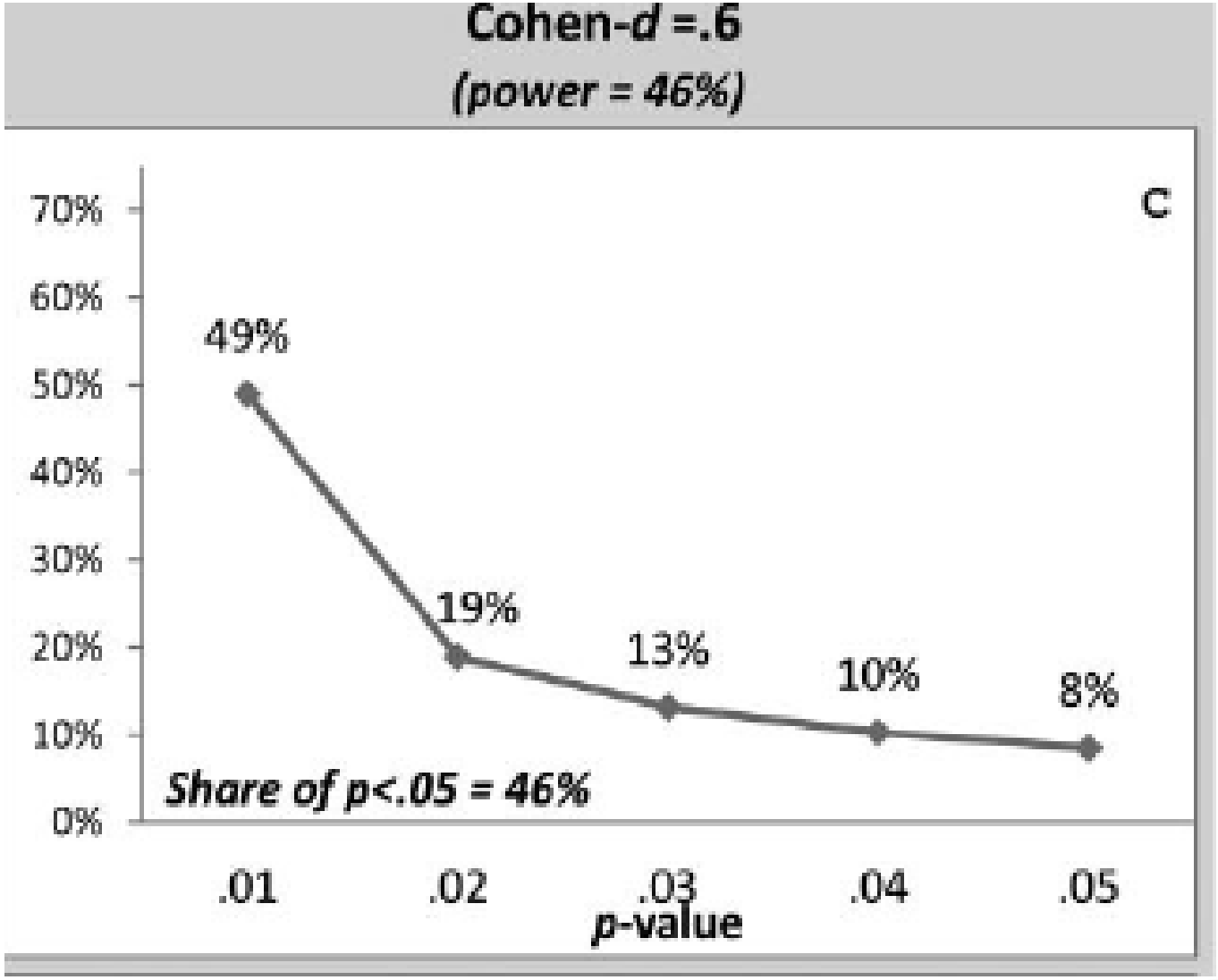
Example: 2-tailed *t*-test ($n_1 = n_2 = 20$)



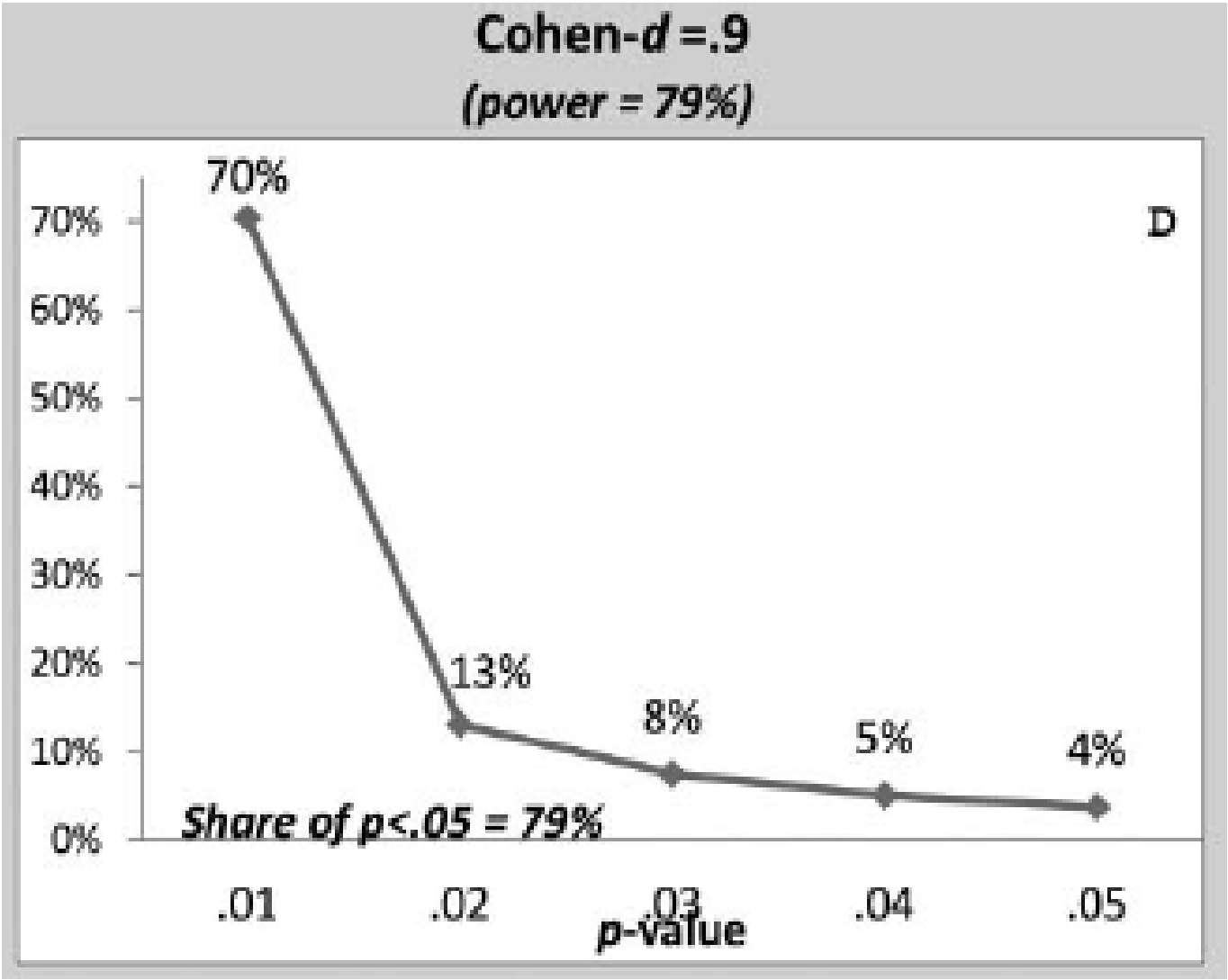
Example: 2-tailed *t*-test ($n_1 = n_2 = 20$)



Example: 2-tailed *t*-test ($n_1 = n_2 = 20$)

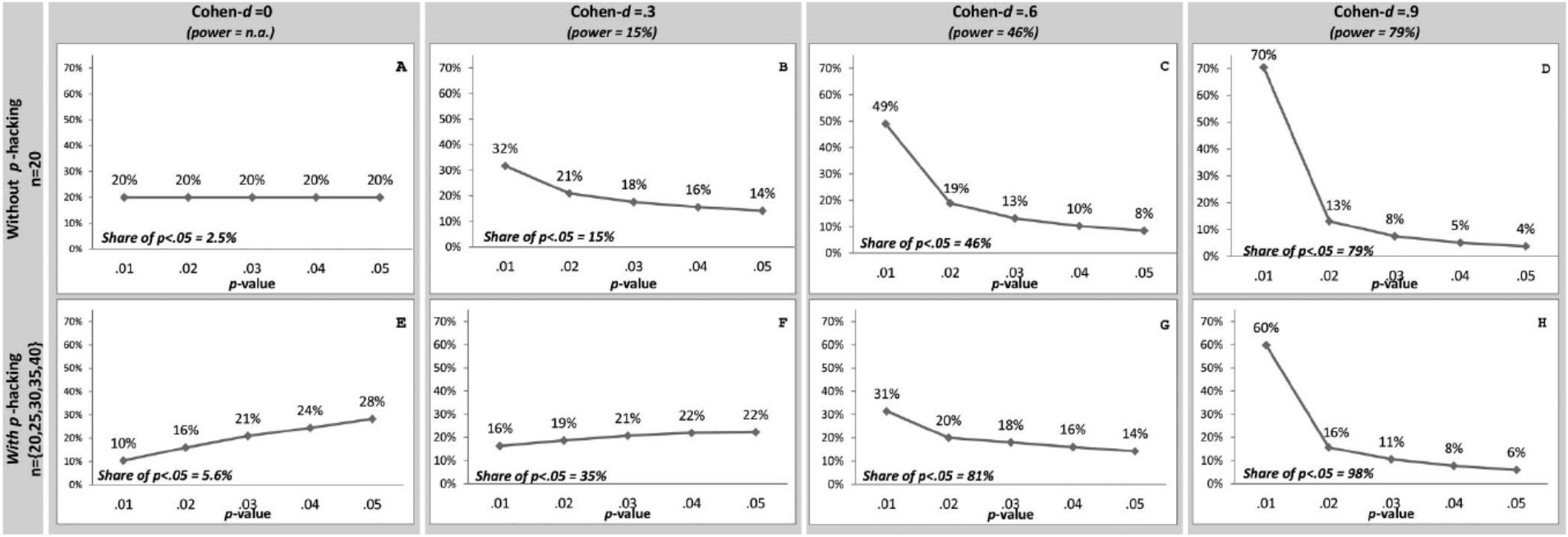


Example: 2-tailed *t*-test ($n_1 = n_2 = 20$)



The role of *p*-hacking

Top: Simulated *t*-test *p*-curves without *p*-hacking
Bottom: Simulated *t*-test *p*-curves with *p*-hacking



Figures E-H: Data peeking with *n*=5-increments for each sample (20-40)

Interpretations suggested by Simonsohn et al. (2014)

Based on the findings that

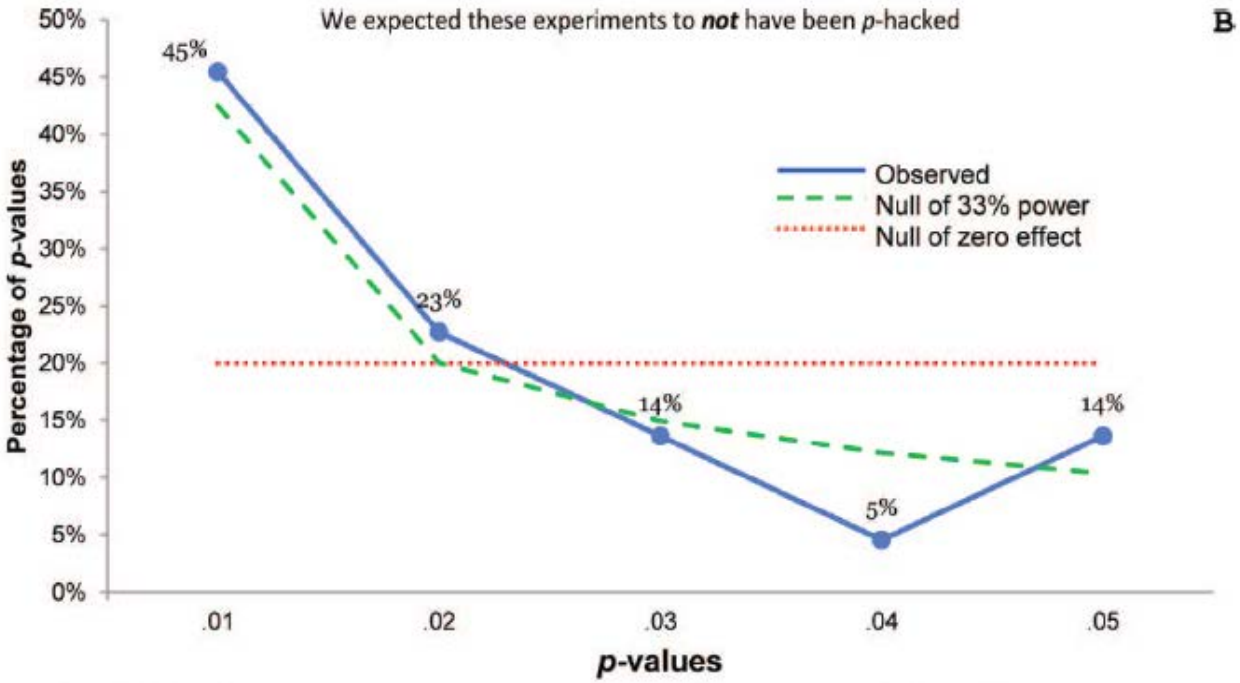
- Null effects → Flat *p*-curves
- True effects → Right-skewed *p*-curves
- *P*-hacked null effects → Left-skewed *p*-curves,

Simonsohn et al. (2014) suggested to reverse the if-then relation, resulting in the following *p*-curve interpretations:

- Flat *p*-curves → No evidential value (null effect)
- Right-skewed *p*-curves → Evidential value (true effect)
- Left-skewed *p*-curves → *p*-hacked null effects

Illustration 1: 20 JPSP studies

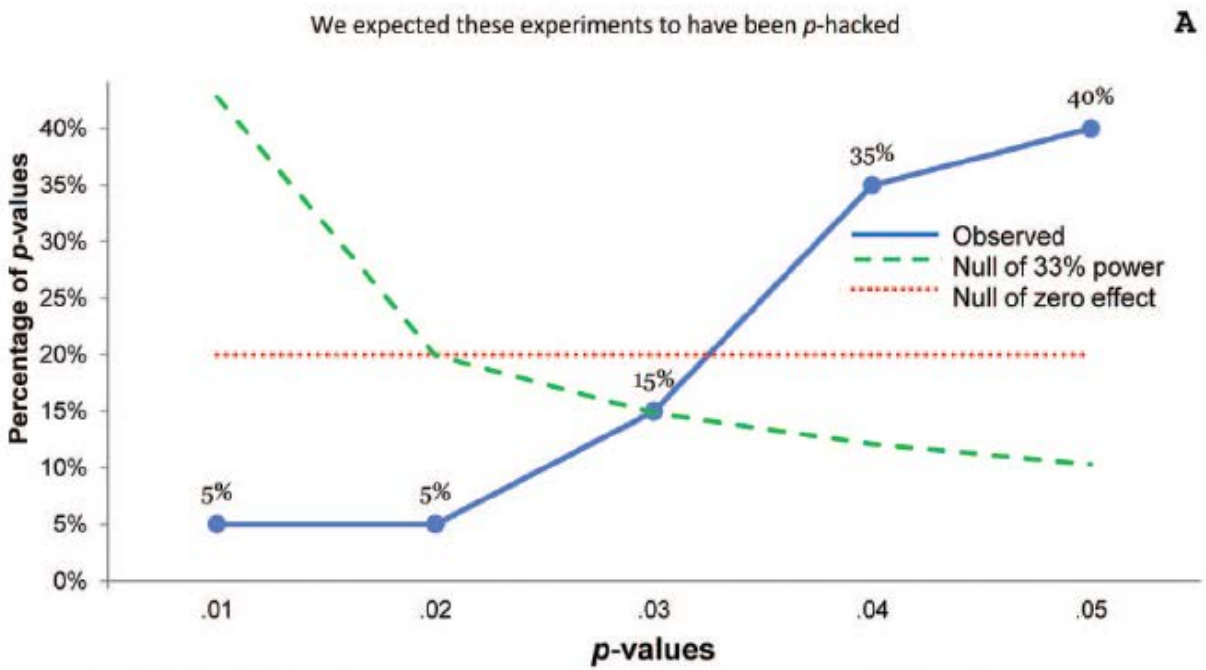
(exclusion criteria: words "covariate", "excluded", "transform")



Statistical Inference	*Results*
1) Studies contain evidential value *(right-skewed)*	$x^2(44)=94.2$, $p<.0001$
2) Studies lack evidential value (flatter than 33%)	$x^2(44)=43.2$, $p=.507$
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(44)=27.2$, $p=.978$

The observed p-curve includes 22 significant p-values, an additional 3 were p>.05
Of those 22 p-values, 16 are p<.025, binomial test for right-skew: p=.026; for left-skew: p=.991.

Illustration 2: 20 JPSP studies
(inclusion criteria: randomized experiment + ANCOVA)



Statistical Inference	*Results*
1) Studies contain evidential value *(right-skewed)*	$x^2(40)=18.3$, $p=.999$
2) Studies lack evidential value (flatter than 33%)	$x^2(40)=82.5$, $p<.0001$
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(40)=58.2$, $p=.031$

The observed p-curve includes 20 significant p-values, an additional 3 were p>.05
Of those 20 p-values, 3 are p<.025, binomial test for right-skew: p>.999; for left-skew: p=.0013

Critical evaluation

- Does the shape of *p*-curve really reveal evidential value or *p*-hacking?
- This includes two questions:
- Does right-skewness imply evidential value?
- Does left-skewness imply p-hacking?
- We will answer these questions one by one.

Does right-skewness imply evidential value? (Ulrich & Miller, 2015)

Journal of Experimental Psychology: General
2015, Vol. 144, No. 6, 1137–1145

0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000086

COMMENT

p-Hacking by Post Hoc Selection With Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)

Rolf Ulrich
University of Tübingen

Jeff Miller
University of Otago

Simonsohn, Nelson, and Simmons (2014) have suggested a novel test to detect *p*-hacking in research, that is, when researchers report excessive rates of "significant effects" that are truly false positives. Although this test is very useful for identifying true effects in some cases, it fails to identify false positives in several situations when researchers conduct multiple statistical tests (e.g., reporting the most significant result). In these cases, *p*-curves are right-skewed, thereby mimicking the existence of real effects even if no effect is actually present.

Right-skewed *p*-curves do not imply evidential value (Ulrich & Miller, 2015)

- Depending on the data analysis strategy, right-skewed *p*-curves may result from null effects (i.e., H_0 holds).
- Examples:
- Try different dependent variables and report the one that produces the smallest *p*-value.
- Combine different dependent variables with “almost” significant outcomes into a composite score.
- Report significant effects only if significance holds for all dependent variables explored.

Simulation results for strategy "Report all significant results if all tests performed are significant"

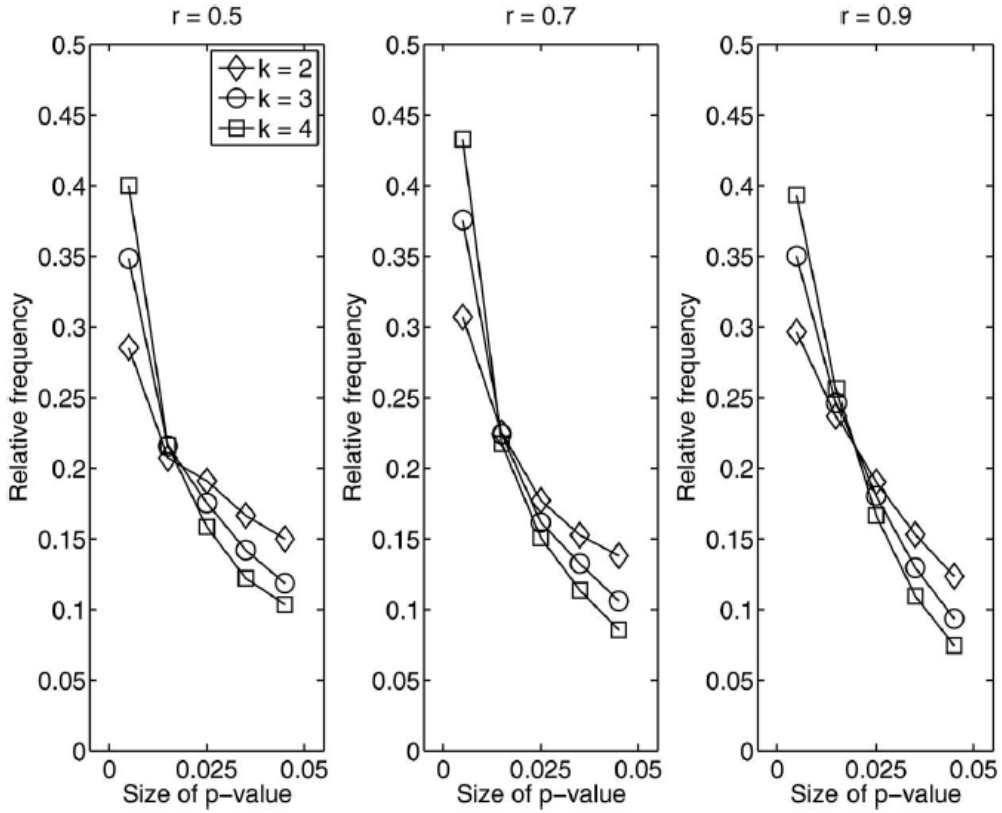


Figure 7. Scenario: "Report all significant results if all k tests are significant." Each curve is based on one million simulated experiments and p values smaller than .05 in these experiments were grouped into five equal sized bins [.00–.01), [.01–.02), [.02–.03), [.03–.04), and [.04–.05). For $r = 0.5$ the relative frequency of all studies reporting exactly k significant tests is .012, .004, and .002 for k equal to 2, 3, and 4, respectively. For $r = 0.7$, these relative frequencies were .019, .011, and .007; for $r = 0.9$ these were .031, .025, and .021, respectively. Note that the range of y-axis had to be increased compared with this range of the previous figures.

Does right-skewness imply evidential value? (Ulrich & Miller, 2018)

Psychological Methods
2018, Vol. 23, No. 3, 546–560

1082-989X/18/$12.00 http://dx.doi.org/10.1037/met0000125

Some Properties of *p*-Curves, With an Application to Gradual Publication Bias

Rolf Ulrich
University of Tübingen

Jeff Miller
University of Otago

Abstract

p-curves provide a useful window for peeking into the file drawer in a way that might reveal *p*-hacking (Simonsohn, Nelson, & Simmons, 2014a). The properties of *p*-curves are commonly investigated by computer simulations. On the basis of these simulations, it has been proposed that the skewness of this curve can be used as a diagnostic tool to decide whether the significant *p* values within a certain domain of research suggest the presence of *p*-hacking or actually demonstrate that there is a true effect. Here we introduce a rigorous mathematical approach that allows the properties of *p*-curves to be examined without simulations. This approach allows the computation of a *p*-curve for any statistic whose sampling distribution is known and thereby allows a thorough evaluation of its properties. For example, it shows under which conditions *p*-curves would exhibit the shape of a monotone decreasing function. In addition, we used weighted distribution functions to analyze how 2 different types of publication bias (i.e., cliff effects and gradual publication bias) influence the shapes of *p*-curves. The results of 2 survey experiments with more than 1,000 participants support the existence of a cliff effect at $p = .05$ and also suggest that researchers tend to be more likely to recommend submission of an article as the level of statistical significance increases beyond this *p* level. This gradual bias produces right-skewed *p*-curves mimicking the existence of real effects even when no such effects are actually present.

Right-skewness may result from gradual publication bias depending on *p*-values

Table 2
Percentage of Participants Answering "Submit the Results Obtained" as a Function of p-Value and Survey Group

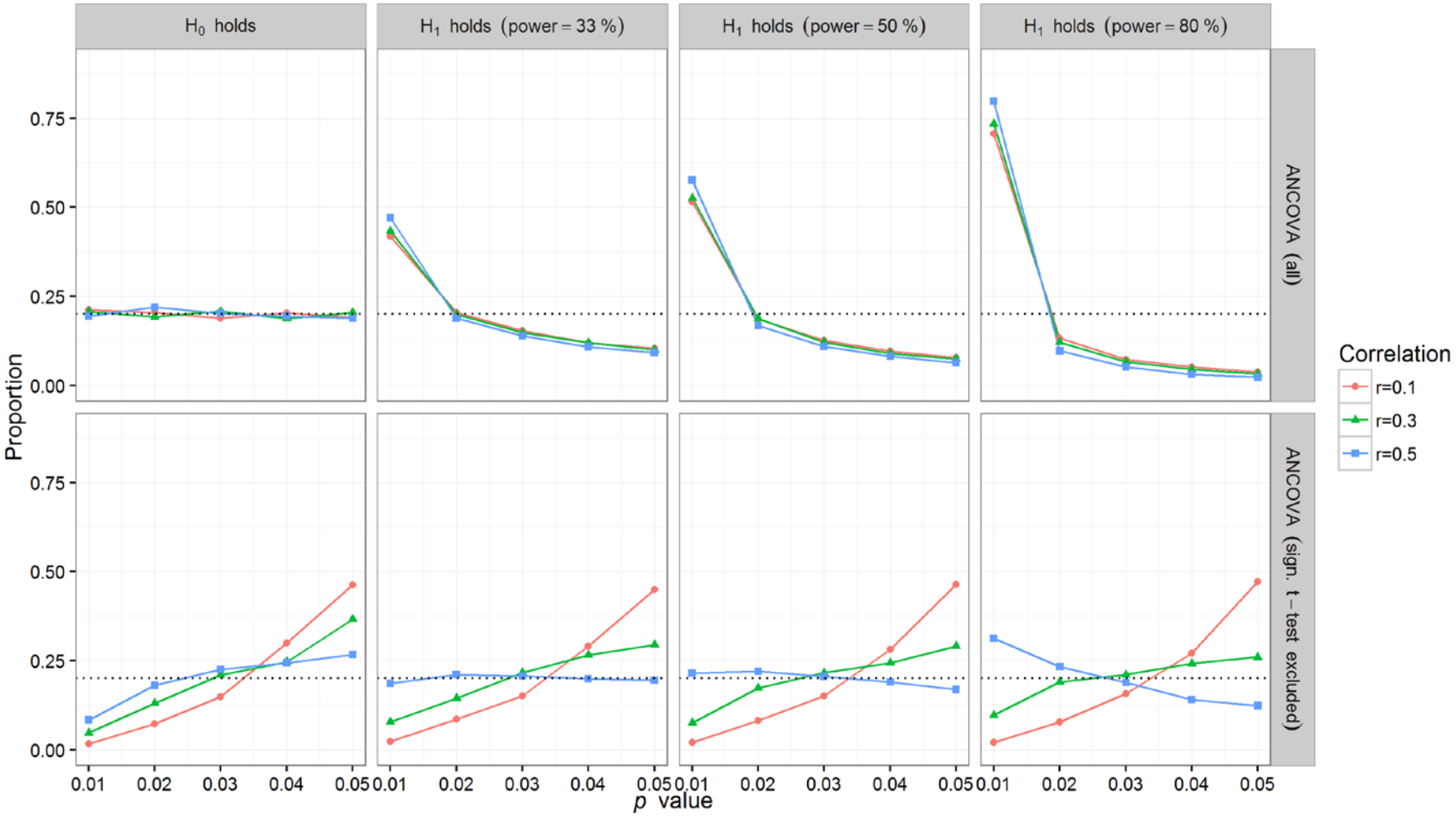
	p-value			
	.0532	.0432	.0232	.0032
German psychologists	24.2	33.8	48.6	52.0
SE	3.5	4.0	4.1	4.1
	p-value			
	.0532	.0432	.0232	.0132
Experimental psychologists	10.2	35.3	36.0	43.5
SE	2.4	3.8	3.9	4.1

Note. *SE* = standard error of estimate.

Does left-skewness imply *p*-hacking?

- Consider Simonsohn et al.'s (2014) ANCOVA example for two randomized groups.
- What is the optimal test? ANCOVA or *t*-test?
- Given substantial effect of covariate, ANCOVA is most powerful:
  - Thus, perform and report ANCOVA only
- Assume researchers know about the advantages of ANCOVA but prefer to report tests that are easier to communicate:
  - Perform ANCOVA first.
  - If significant, try a *t*-test (because it requires less assumptions and is thus easier to communicate).
  - If significant, report the *t*-test only.
- Consequence:
  - Significant *t*-tests are excluded from ANCOVA *p*-curve.

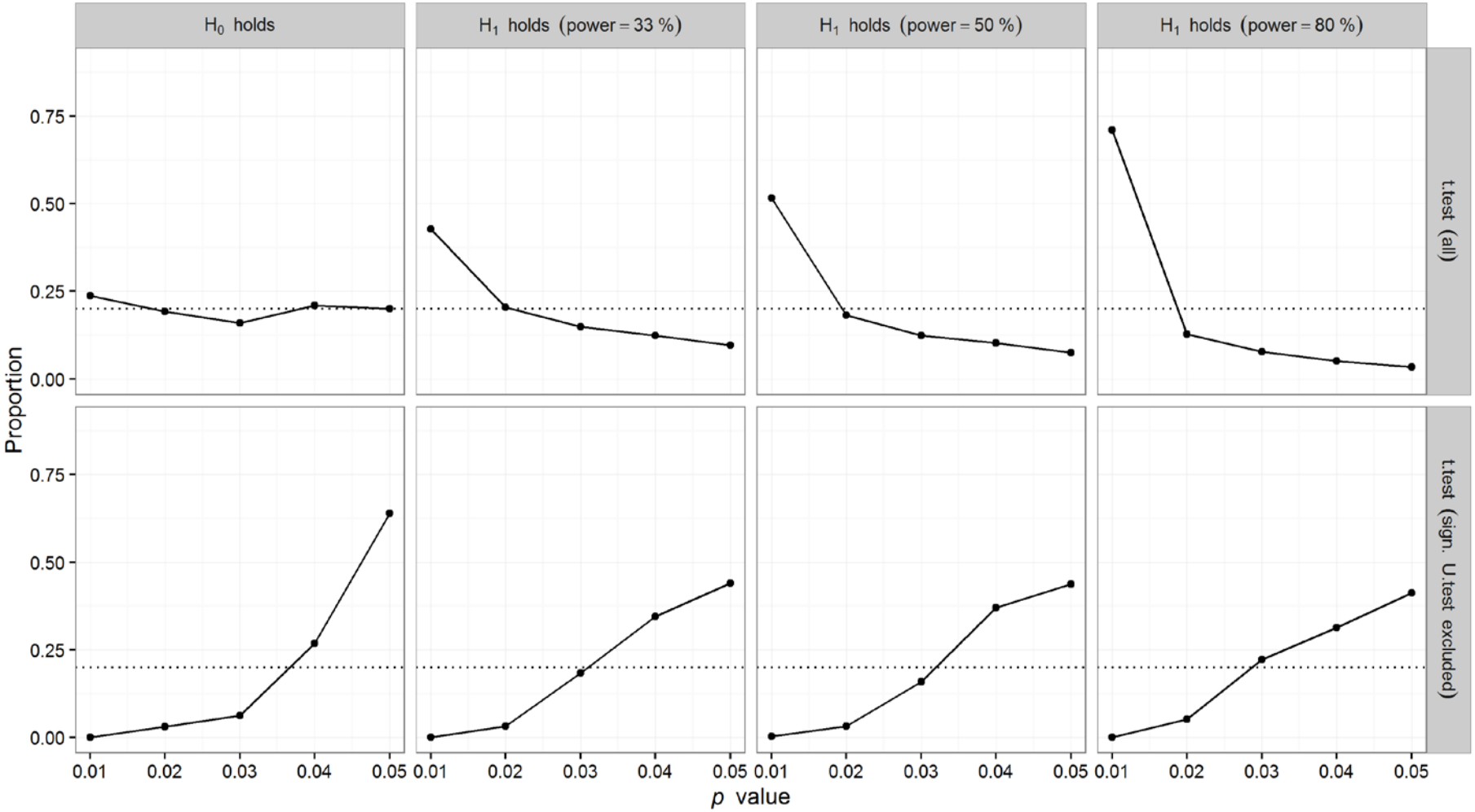
Top: Unselected ANCOVA *p*-curve
Bottom: ANCOVA *p*-curve given preference for *t*-tests



Is the problem specific for ANCOVAs?

- The results obtained for ANCOVA *p*-curves generalize to other test scenarios.
- General rule:
  - Whenever we switch from a powerful test of a hypothesis to a less powerful test and report only the last one if significant, the *p*-curve for the more powerful test tends to be flatter or even left-skewed.
- Another example is the *t*-test *p*-curve, given preference to report *U*-tests (quite common in the neuro- and biosciences)
  - Try a two-groups *t*-test first.
  - If significant, try *U*-test (because it is more robust and thus easier to communicate, at least in some scientific communities)
  - If significant, report the *U*-test result only.

Top: Unselected *t*-test *p*-curve
Bottom: *t*-test *p*-curve given preference for *U*-tests



Conclusions

- Right-skewness of *p*-curves does not imply evidential value.
- Left-skewness of *p*-curves does not imply *p*-hacking.
- Implications for the *p*-curve tool:
  - *P*-curves should be interpreted with caution.
  - *P*-curve analyses may nevertheless be useful.
  - However, explanations for skewed *p*-curves require VALIDATION
    - ANCOVA-example: Check whether covariates were selected ad hoc or theory-driven based on prior research.
    - *U*-test example: Were there statistical reasons to prefer the *U*-test over the *t*-test?
- Implications for data analysis strategies
  - Perform and report the “optimal” test only!!
  - The optimal test should be defined a priori (i.e., registered report).
  - In any case, report all analyses performed (see APA guidelines)!!

Thank you

for your attention!

Figure S2. Expected *p*-curves from sets of studies reporting ANCOVA where some researchers, 'choosers,' are excluded because they report ANOVA instead if it is $p<.05$.