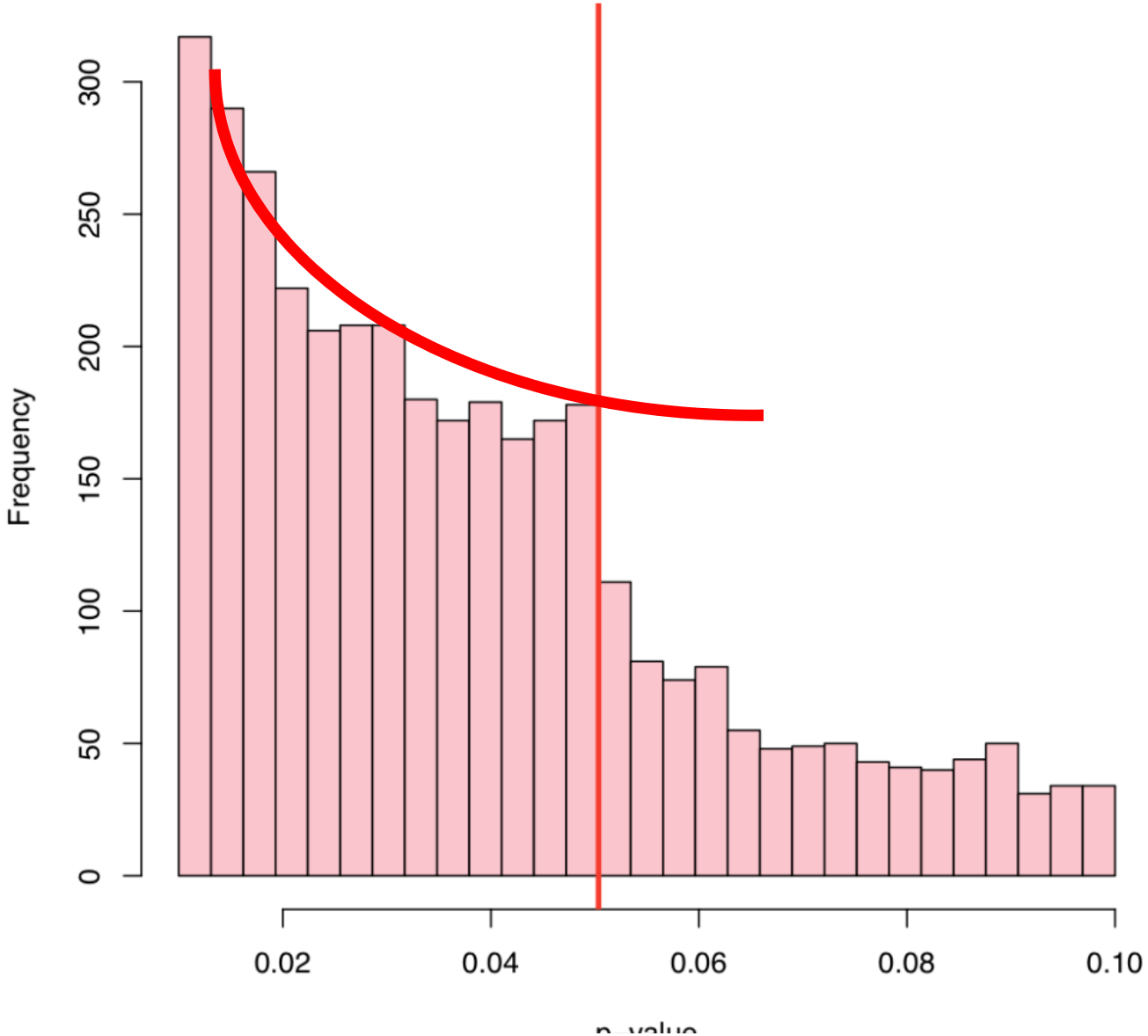
Towards a reliable and cumulative psychological science

Daniel Lakens
School of Innovation Sciences
Eindhoven University of Technology

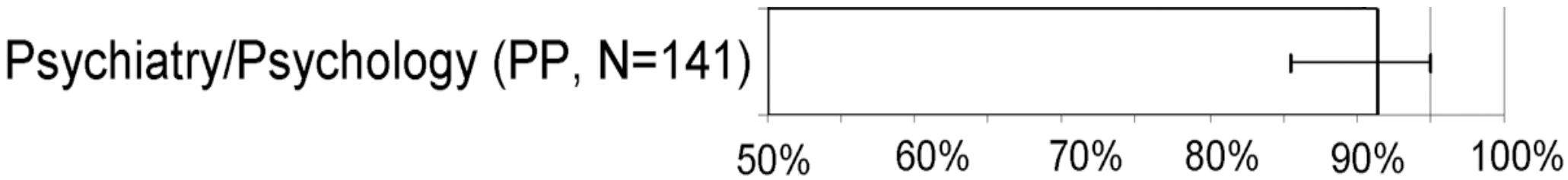
A little bit about the problem, and then on to some solutions.

One of the bigger problems of *p*-values is their use as a threshold to publish.


Frequency
0
50
100
150
200
250
300
0.02
0.04
0.06
0.08
0.10

And this is where we put the non-significant results.

Most published findings confirm the hypothesis (Fanelli, 2010)



Non-significant studies should be expected:

$0.8 \times 0.8 \times 0.8 \times 0.8 = 0.41$

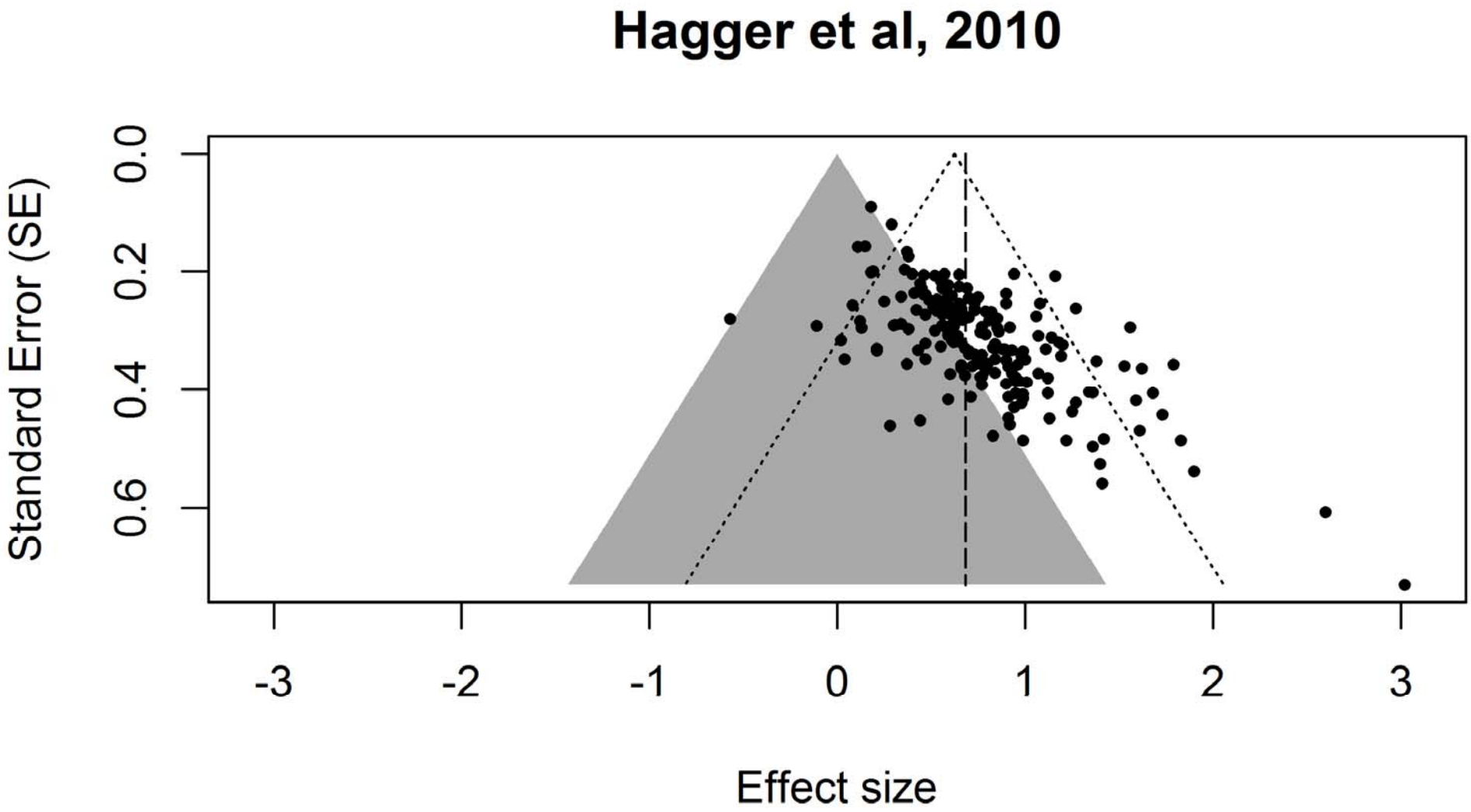
As long as a research area doesn’t share all results, it’s not a quantitative science.

Promoting excellence in
parapsychological research and education

Koestler Parapsychology Unit

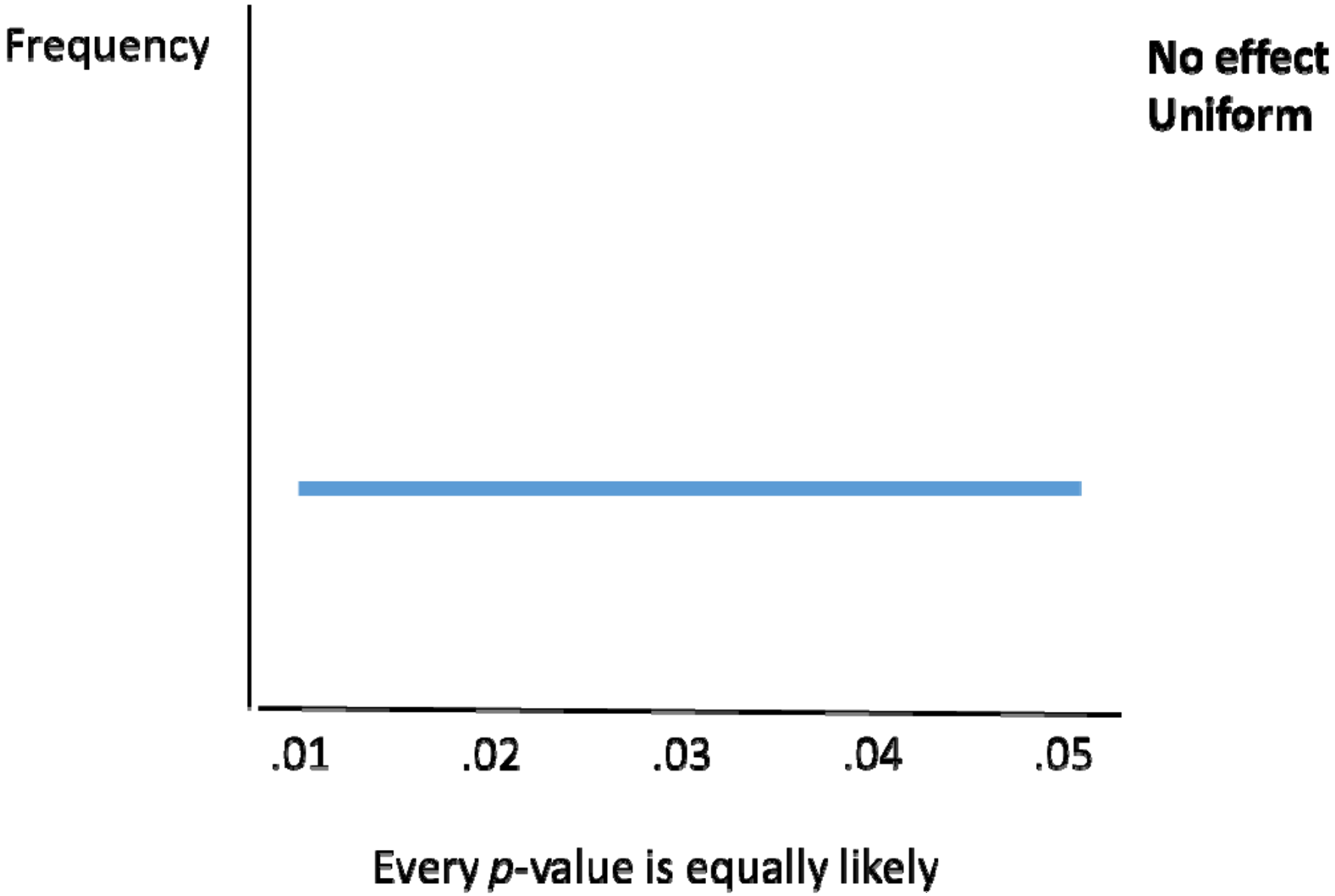
There can be 200 published studies with $p < 0.05$, but no true effect.

Publication bias can not be **corrected**, but it can be **detected**.

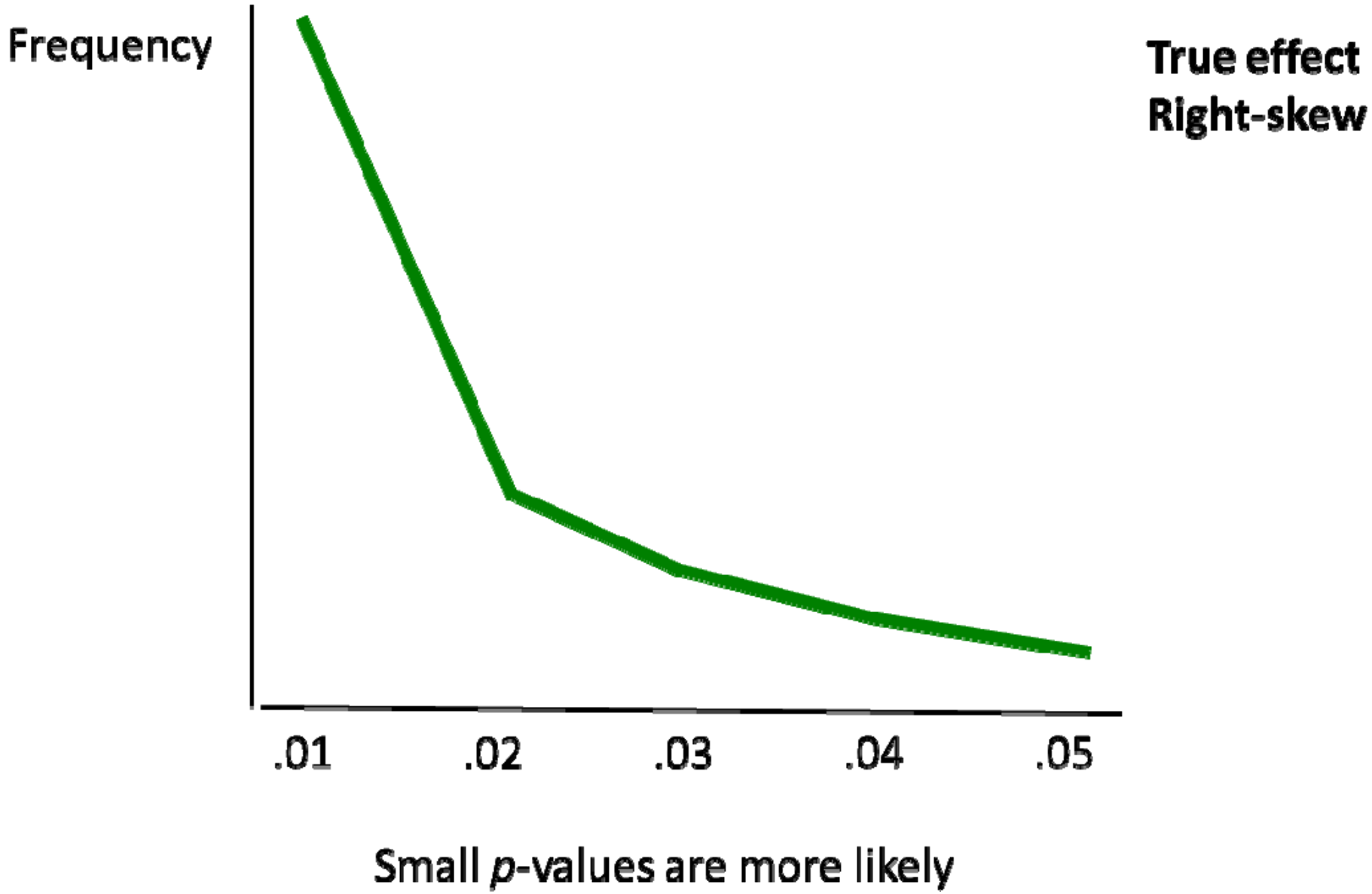

Hagger et al, 2010
Standard Error (SE)
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Effect size

P-curve Analysis

What do *p*-values look like from 100 studies with an effect size of 0?


Frequency
No effect
Uniform
.01
.02
.03
.04
.05
Every *p*-value is equally likely

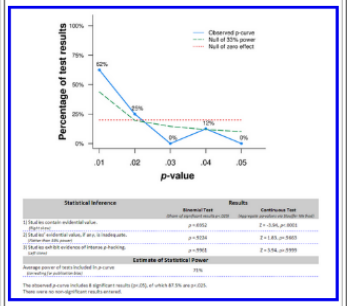
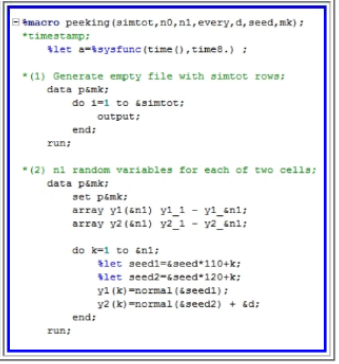
What do *p*-values look like from 100 studies with a true effect?


Frequency
True effect
Right-skew
.01
.02
.03
.04
.05
Small *p*-values are more likely

P-curve analysis: Test whether a **set** of *p*-values has evidential value.

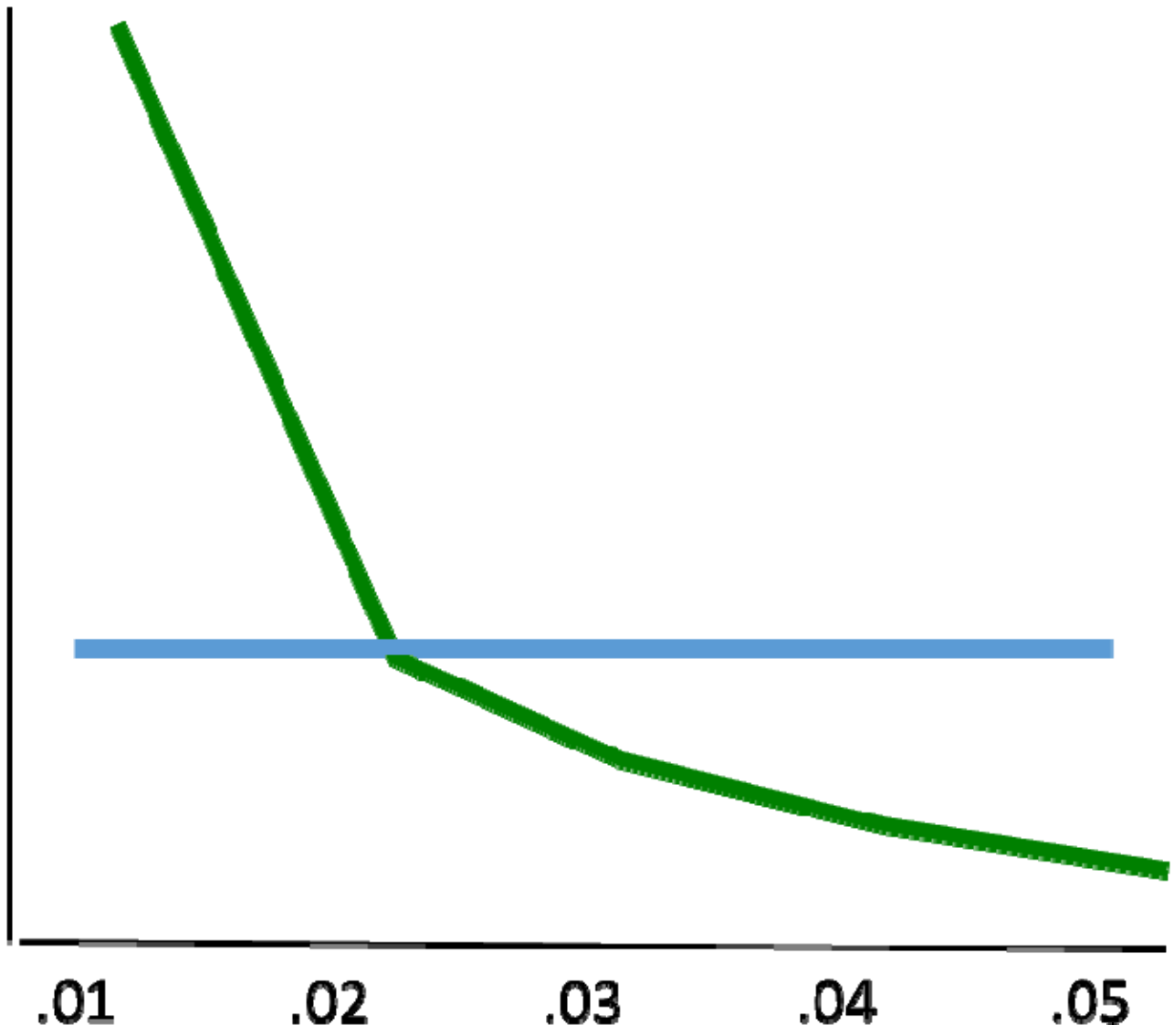
Key to the **file-drawer**: Test is only performed on $p < 0.05$!

P-curve.com

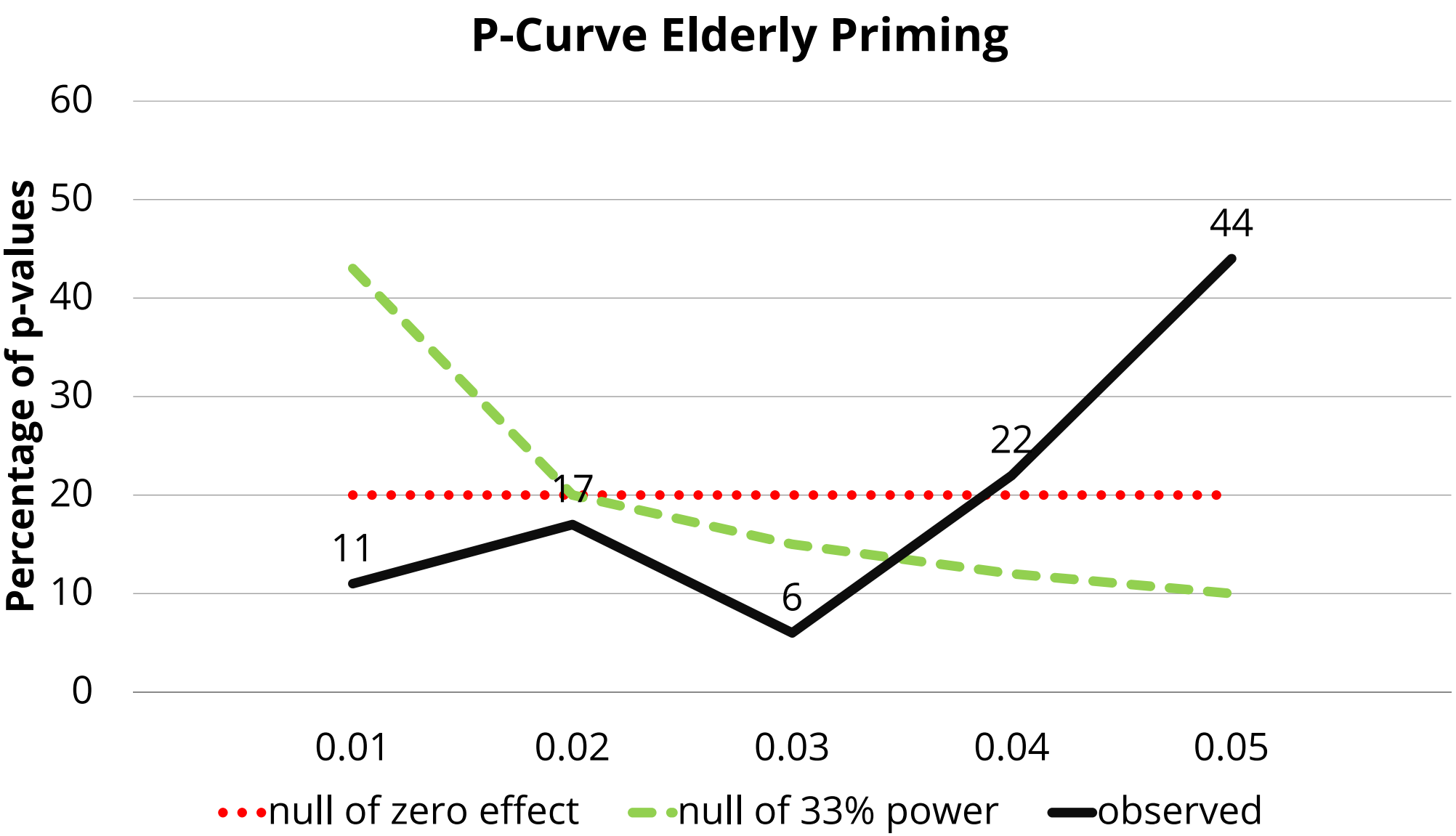
Paper 1 Evidential Value	Paper 2 Effect size	Paper 3 'Better P-curves' (robustness)	The online app 4.0	The User Guide	Supp Materials
					

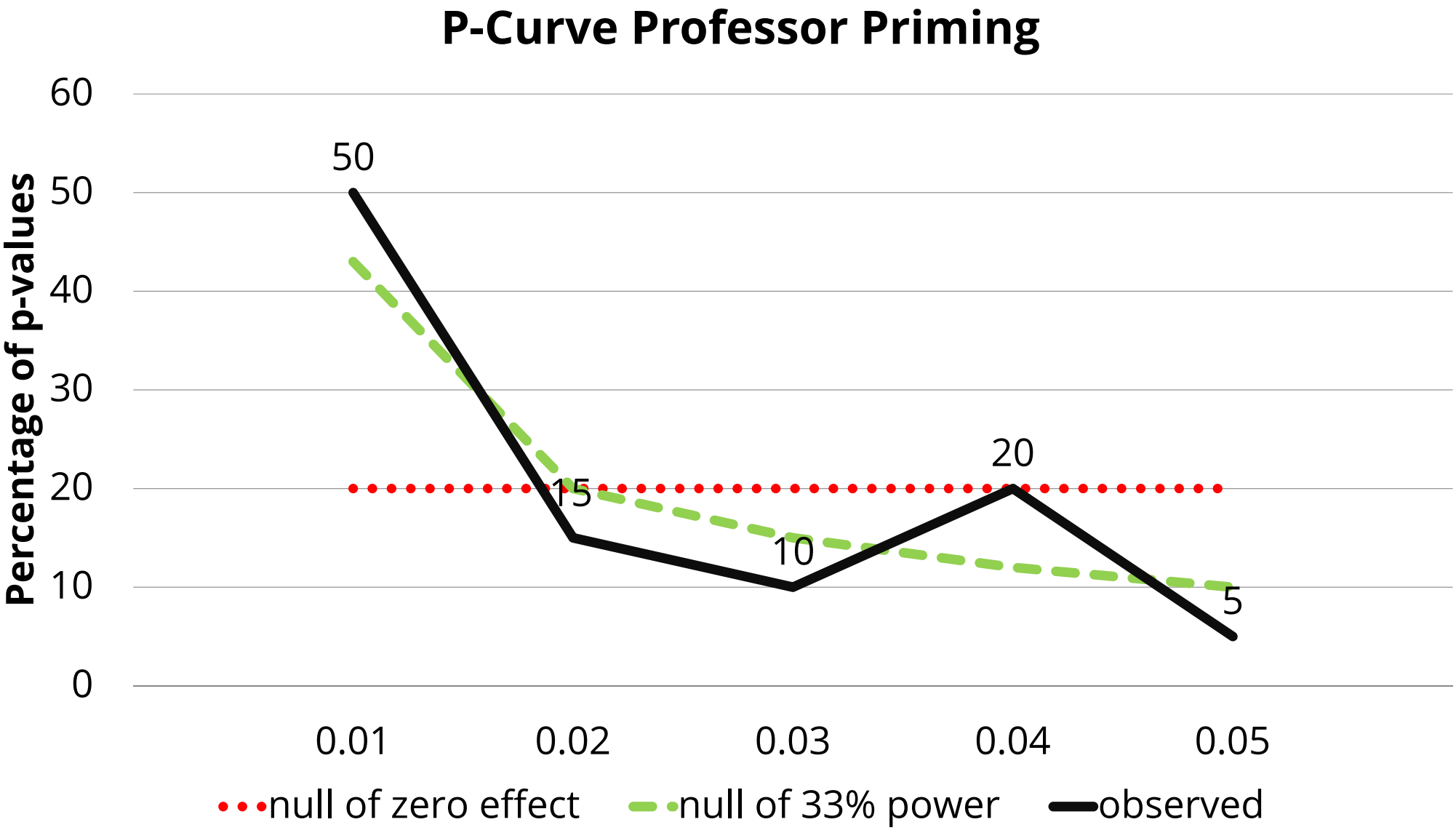

```
%macro peeking(simtot,n0,n1,every,d,seed,mk);
*timestamp;
    %let a=%sysfunc(time(),time8.) ;

*(1) Generate empty file with simtot rows;
    data p&mk;
        do i=1 to &simtot;
            output;
        end;
    run;

*(2) n1 random variables for each of two cells;
    data p&mk;
        set p&mk;
        array y1(&n1) y1_1 - y1_&n1;
        array y2(&n1) y2_1 - y2_&n1;

        do k=1 to &n1;
            %let seed1=&seed*110+k;
            %let seed2=&seed*120+k;
            y1(k)=normal(&seed1);
            y2(k)=normal(&seed2) + &d;
        end;
    run;
```

Does the *p*-value distribution look like one with or one without an effect?


.01
.02
.03
.04
.05

Looking at Elderly Priming and Professor Priming.


P-Curve Elderly Priming
Percentage of p-values
60
50
40
30
20
10
0
0.01
0.02
0.03
0.04
0.05
11
17
6
22
44
null of zero effect
null of 33% power
observed


P-Curve Professor Priming
Percentage of p-values
60
50
40
30
20
10
0
50
15
10
20
5
0.01
0.02
0.03
0.04
0.05
null of zero effect
null of 33% power
observed

You can use this technique for small sets of *p*-values (with care!).

Science AAAS

Home News Journals Topics Careers

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

SHARE

REPORT

Reading Literary Fiction Improves Theory of Mind

David Comer Kidd*, Emanuele Castano*

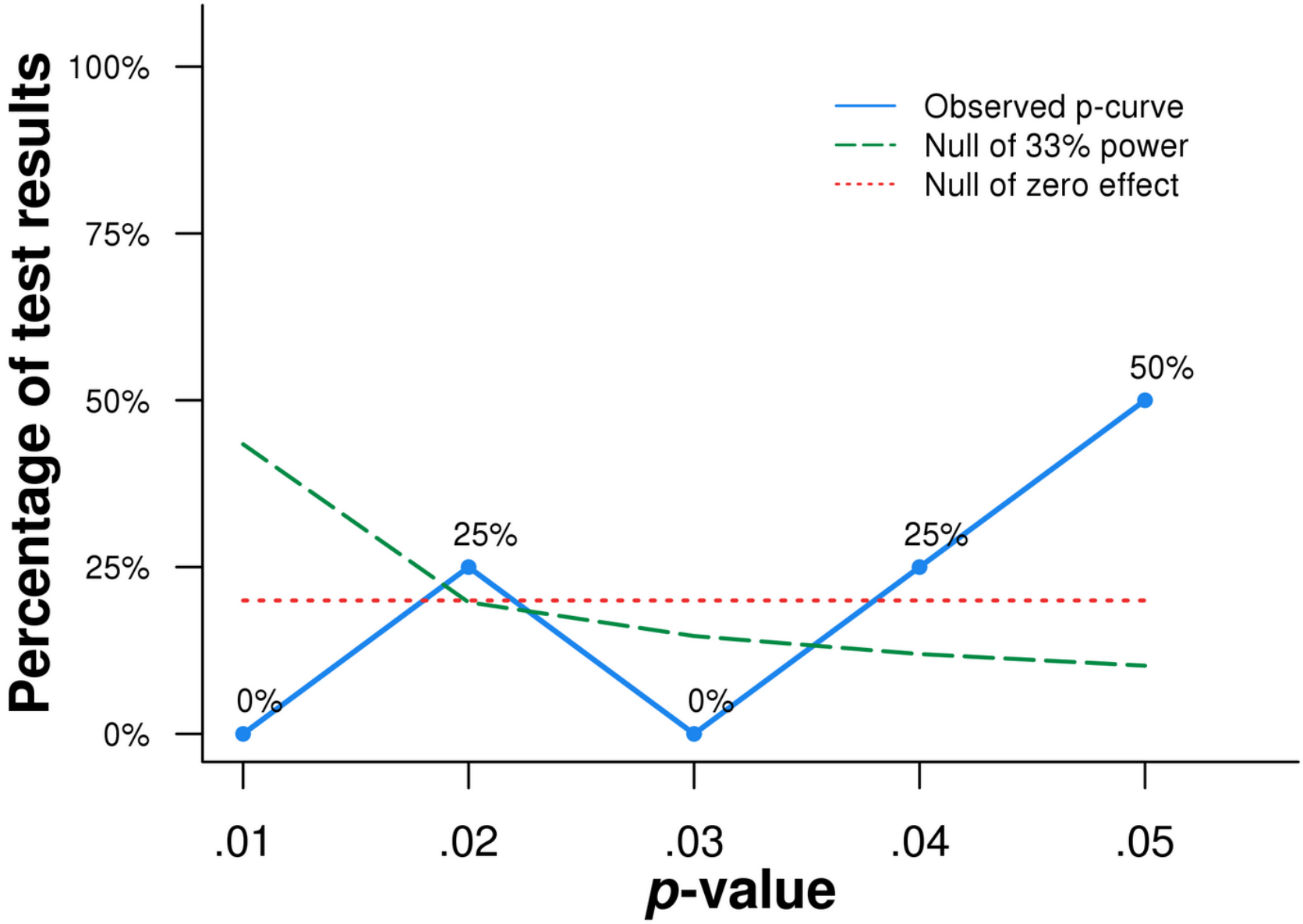
+ Author Affiliations

*Corresponding author. E-mail: kiddd305@newschool.edu (D.C.K.); castanoe@newschool.edu (E.C.)

Science 18 Oct 2013:
Vol. 342, Issue 6156, pp. 377-380

Table 1. RMET and DANVA2-AF analyses.

Experiment	Independent variable	Test	*P*	ω_p^2
Exp. 1 RMET	Condition	$F_{1,82} = 6.40$	0.01	0.05
	Author Recognition Test	$\beta = 0.36$	0.0003	0.13
	Author Recognition Test x Condition	$F_{1,82} = 1.06$	0.30	0.00
Exp. 2 DANVA2-AF	Condition	$F_{2,108} = 2.57$	0.08	0.02
	Author Recognition Test	$\beta = -0.16$	0.08	0.01
	Author Recognition Test x Condition	$F_{2,108} = 1.17$	0.31	0.00
Exp. 3 RMET	Condition	$F_{1,65} = 4.07$	0.04	0.04
	Author Recognition Test	$\beta = -0.01$	0.90	−0.01
	Author Recognition Test x Condition	$F_{1,65} = 0.01$	0.90	−0.01
Exp. 4 RMET	Condition	$F_{1,68} = 4.39$	0.04	0.04
	Author Recognition Test	$\beta = 0.39$	<0.001	0.15
	Author Recognition Test x Condition	$F_{1,68} = 1.50$	0.22	0.00
Exp. 5 RMET	Condition	$F_{2,352} = 3.10$	0.04	0.01
	Author Recognition Test	$\beta = 0.28$	<0.001	0.07
	Author Recognition Test x Condition	$F_{2,352} = 1.37$	0.25	0.00


Percentage of test results
100%
75%
50%
25%
0%
Observed p-curve
Null of 33% power
Null of zero effect
50%
25%
25%
0%
0%
.01
.02
.03
.04
.05
p-value

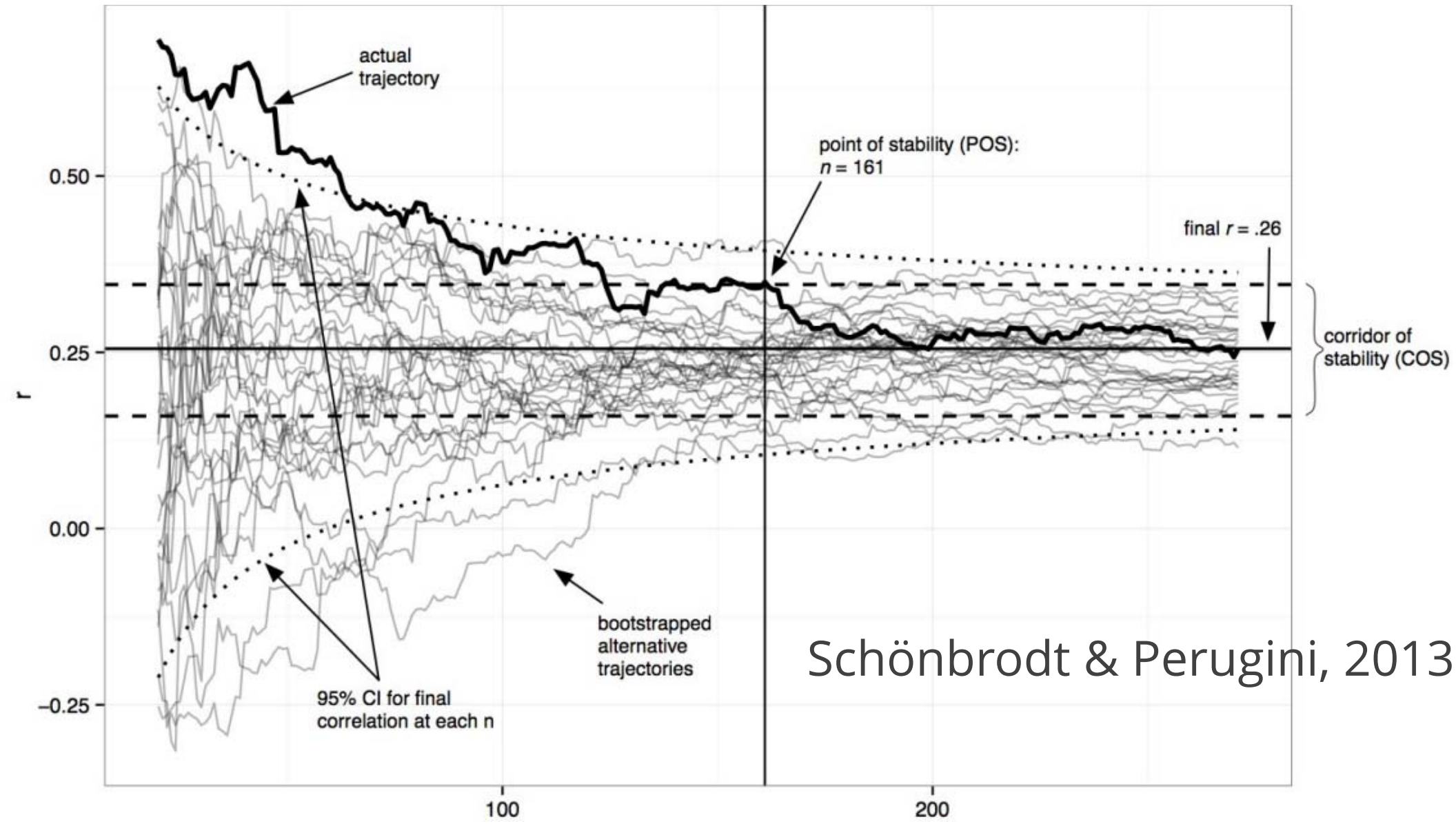
A theory might be true, the data just don’t provide evidence for it.

P-curve tells you if significant *p*-values look more like a **true** or **null** effect.

Sample Size Justification

How do you determine the sample size for a new study?

Small samples have large variation, more Type 2 errors, and inaccurate estimates.



Schönbrodt & Perugini, 2013

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Psychological Bulletin
1989, Vol. 105, No. 2, 309–316

0033-2909/89/$00.75

Do Studies of Statistical Power Have an Effect on the Power of Studies?

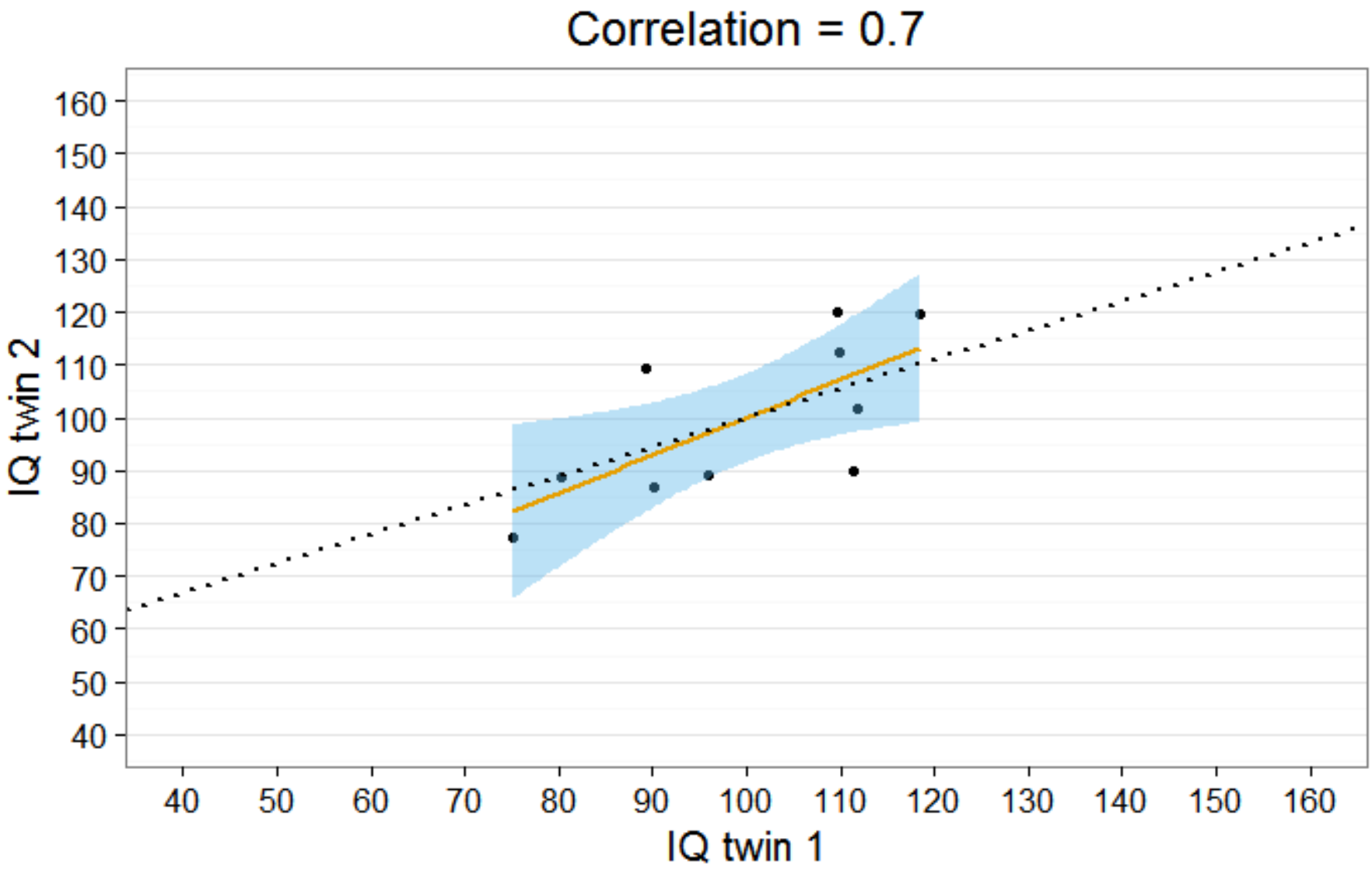
Peter Sedlmeier and Gerd Gigerenzer
University of Konstanz, Federal Republic of Germany

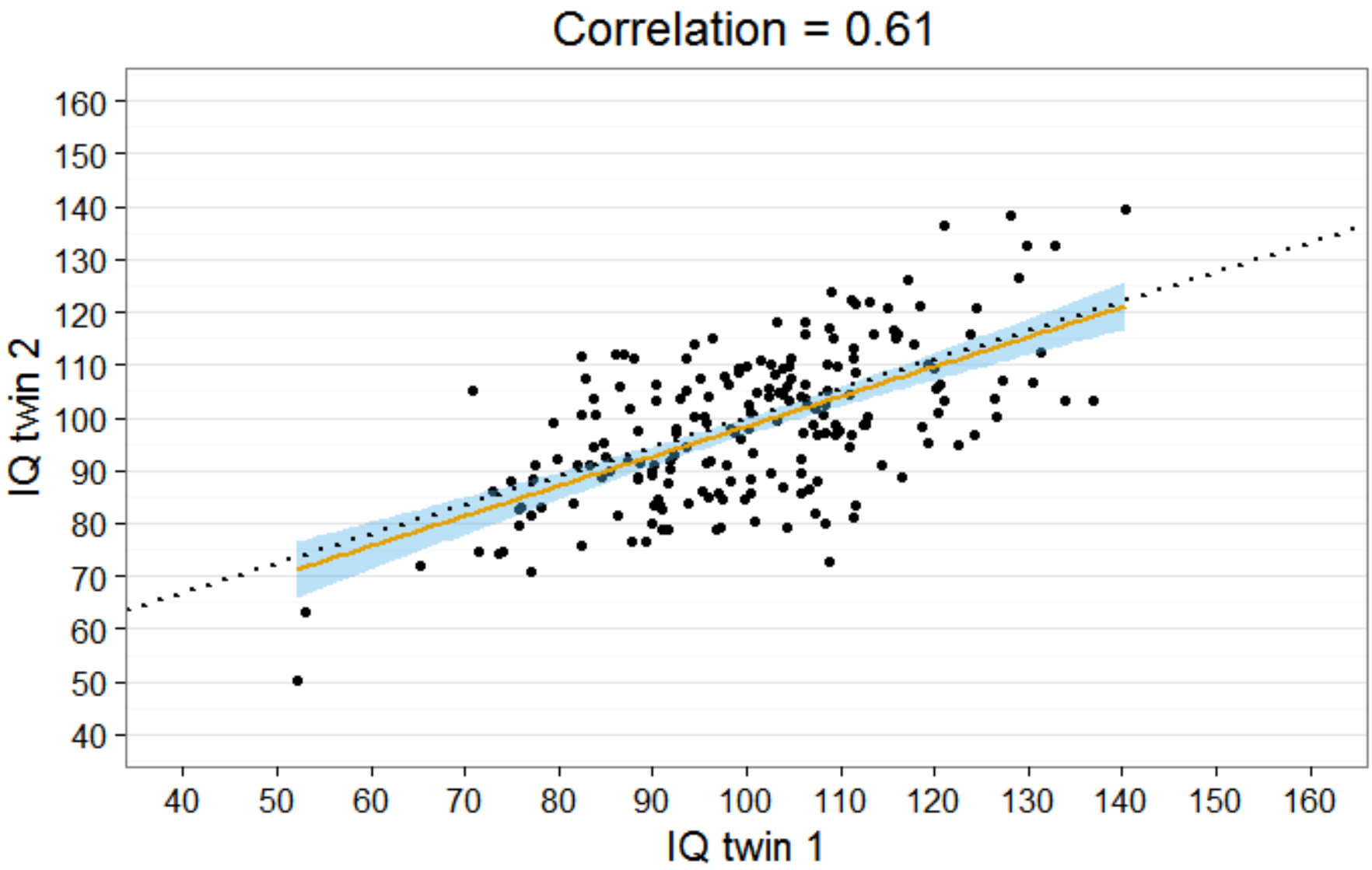
Studies in psychology often have low power. Estimates average around 50%.

Cohen, 1962; Fraley & Vazire, 2014

One reason for low power is that people use heuristics to plan their sample size.

You need to justify the sample size of a study. What goal do you want to achieve?

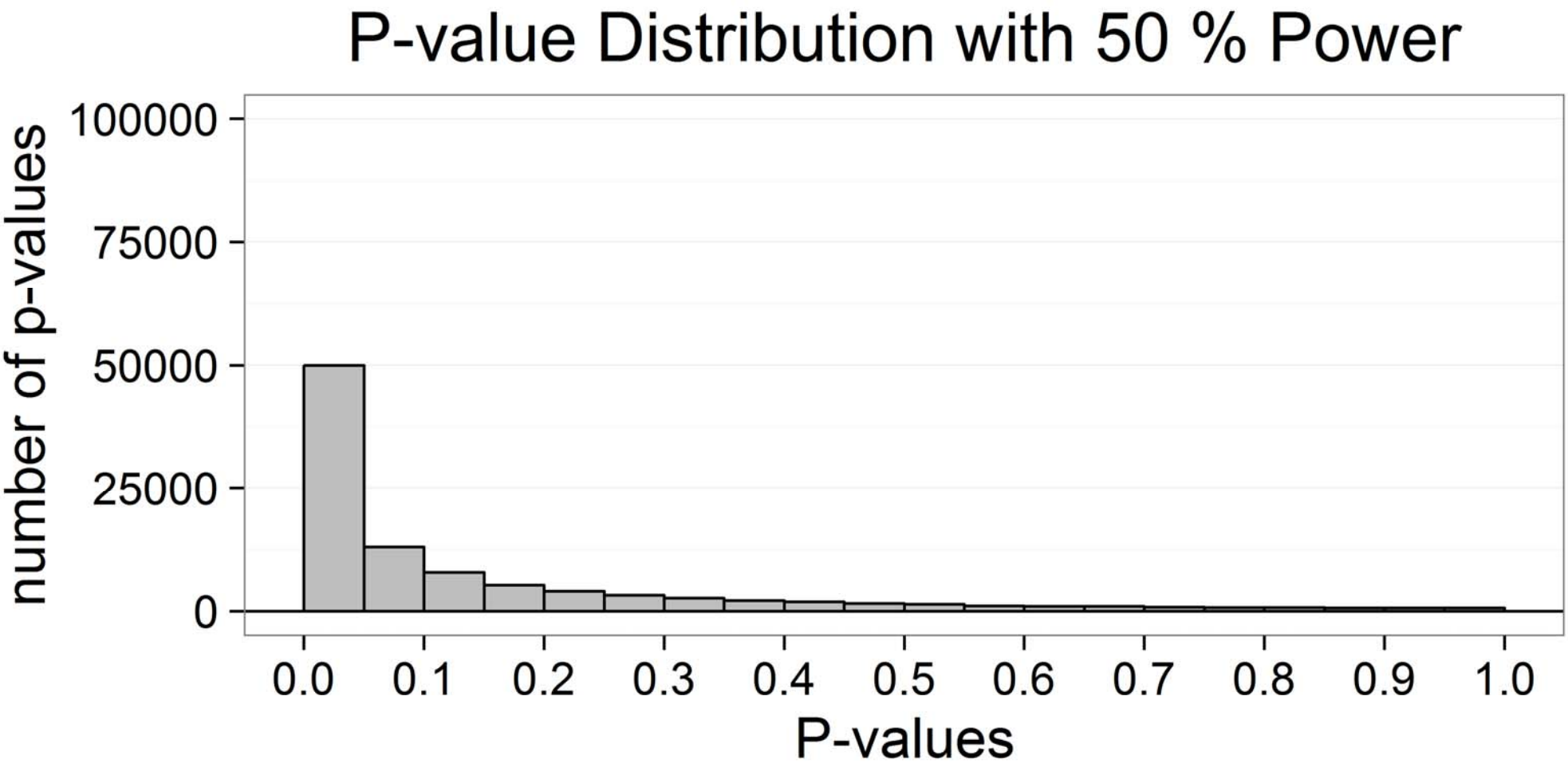

Correlation = 0.7
IQ twin 2
IQ twin 1
160
150
140
130
120
110
100
90
80
70
60
50
40
40
50
60
70
80
90
100
110
120
130
140
150
160

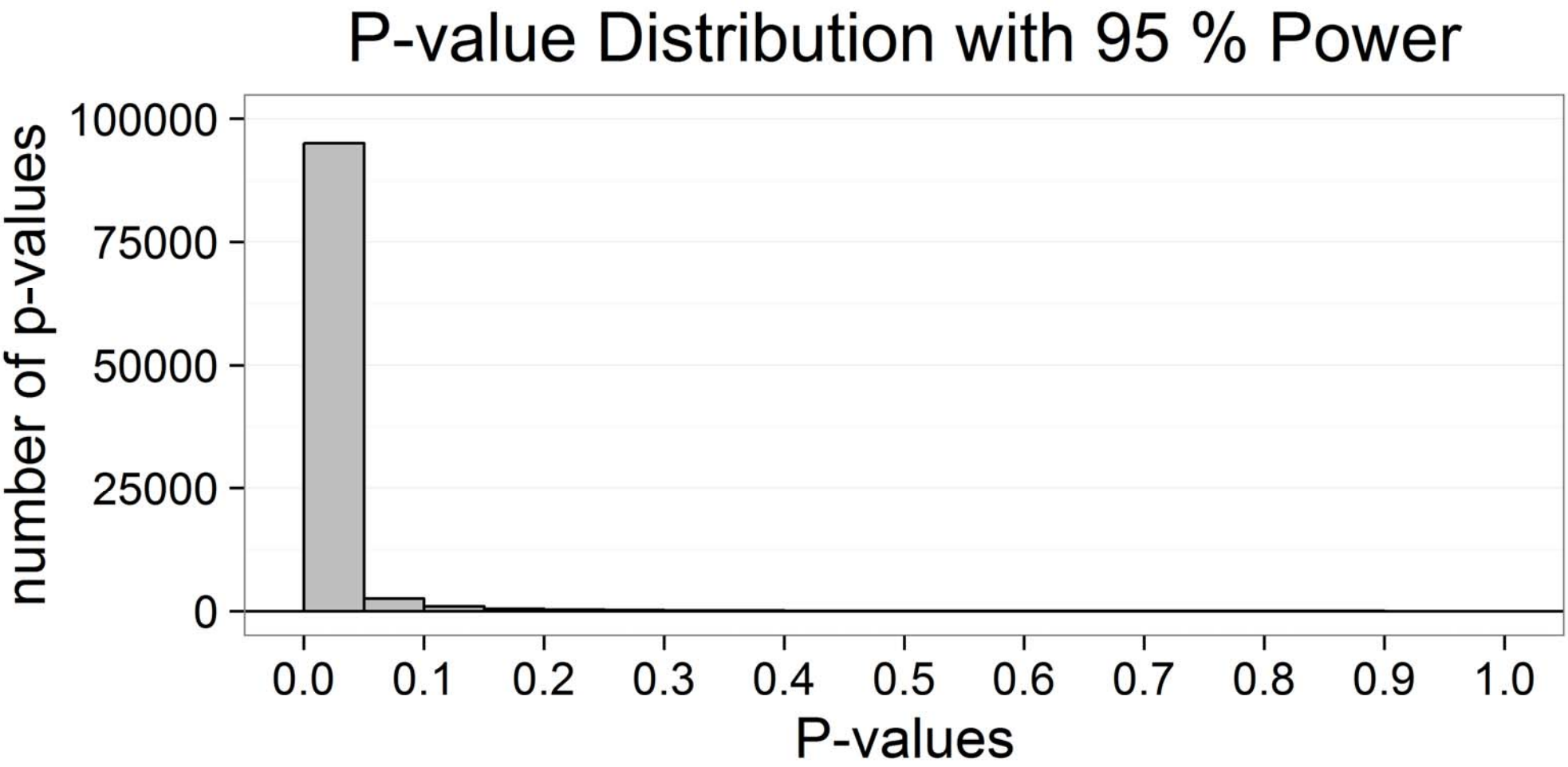

Correlation = 0.61
IQ twin 2
IQ twin 1
160
150
140
130
120
110
100
90
80
70
60
50
40
40
50
60
70
80
90
100
110
120
130
140
150
160

Planning for accuracy

Select a sample size based on the width of the confidence interval

Maxwell, Kelley, & Rausch, 2008


P-value Distribution with 50 % Power
number of p-values
100000
75000
50000
25000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
P-values

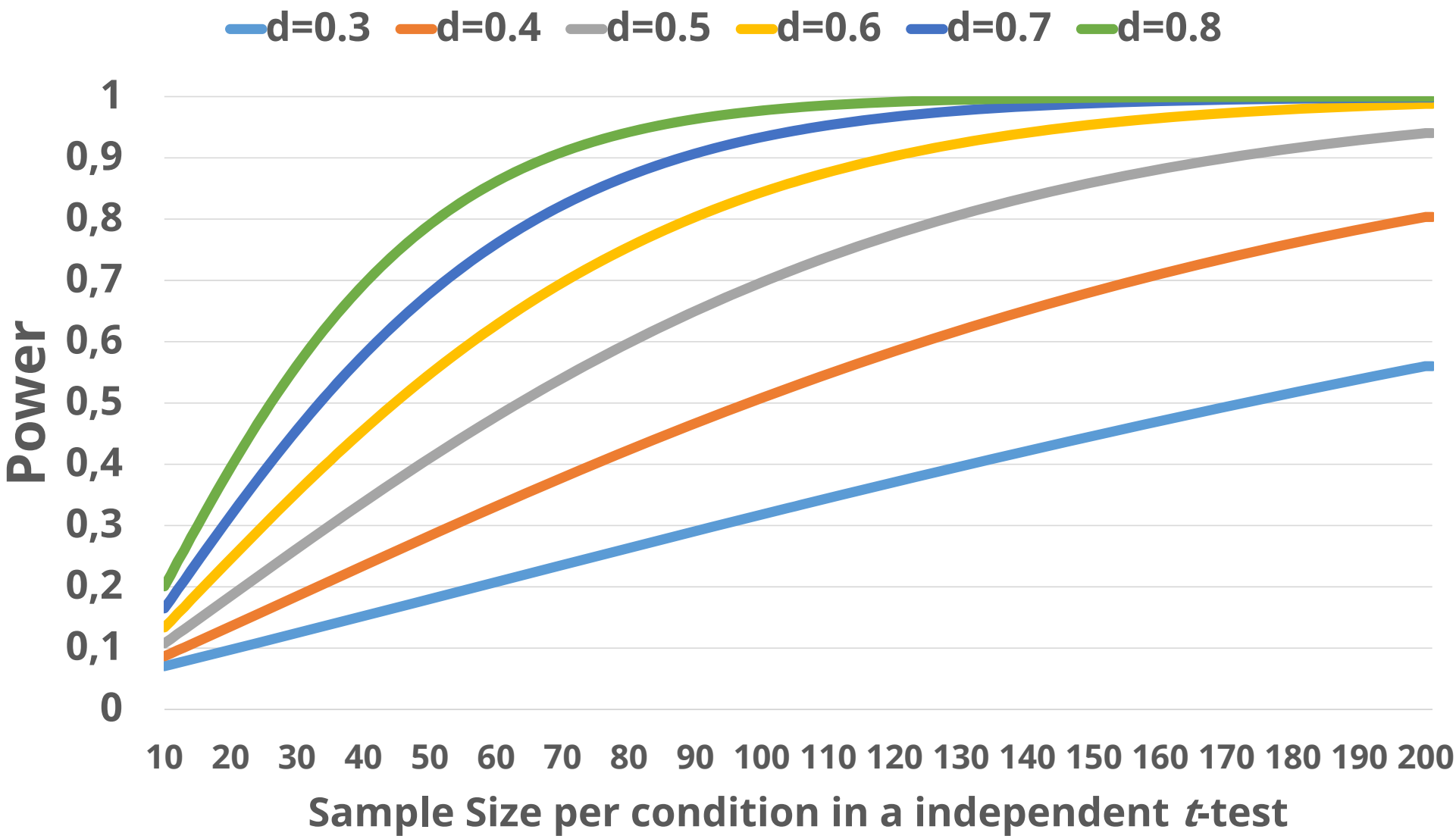

P-value Distribution with 95 % Power
number of p-values
100000
75000
50000
25000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
P-values

Planning for power

Select a sample size based on probability of finding $p < 0.05$.

Planning for power

Yesterday’s "Sample size in psychology research” workshop


d=0.3
d=0.4
d=0.5
d=0.6
d=0.7
d=0.8
Power
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
Sample Size per condition in a independent t-test

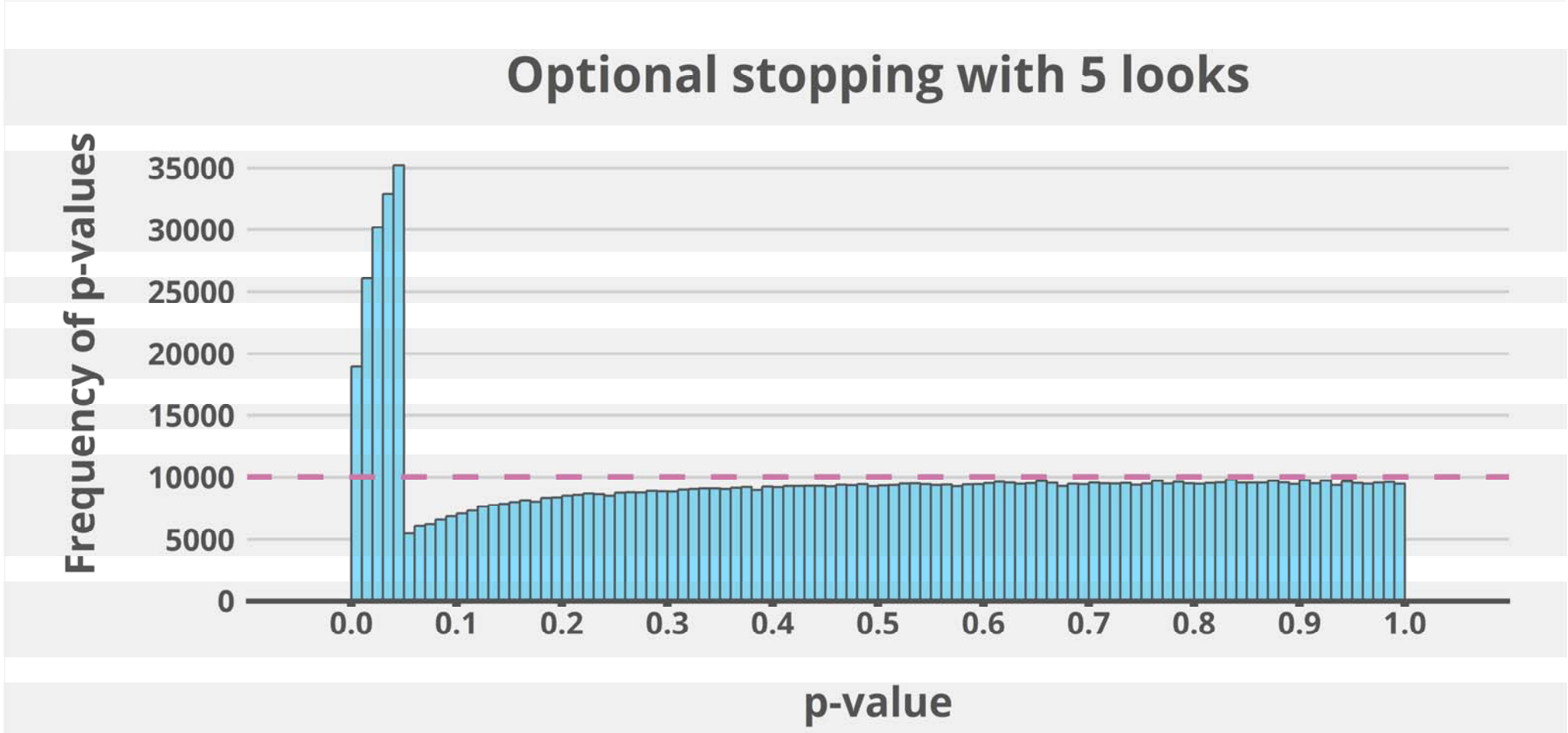
Take care when using effect sizes from the literature. Publication bias inflates effects.

If effect sizes are uncertain **sequential analyses** let you look at data as it comes in.

Optional stopping: Collecting data until $p < 0.05$ inflates the Type 1 error.

A user of NHST could **always** obtain a significant result through optional stopping.

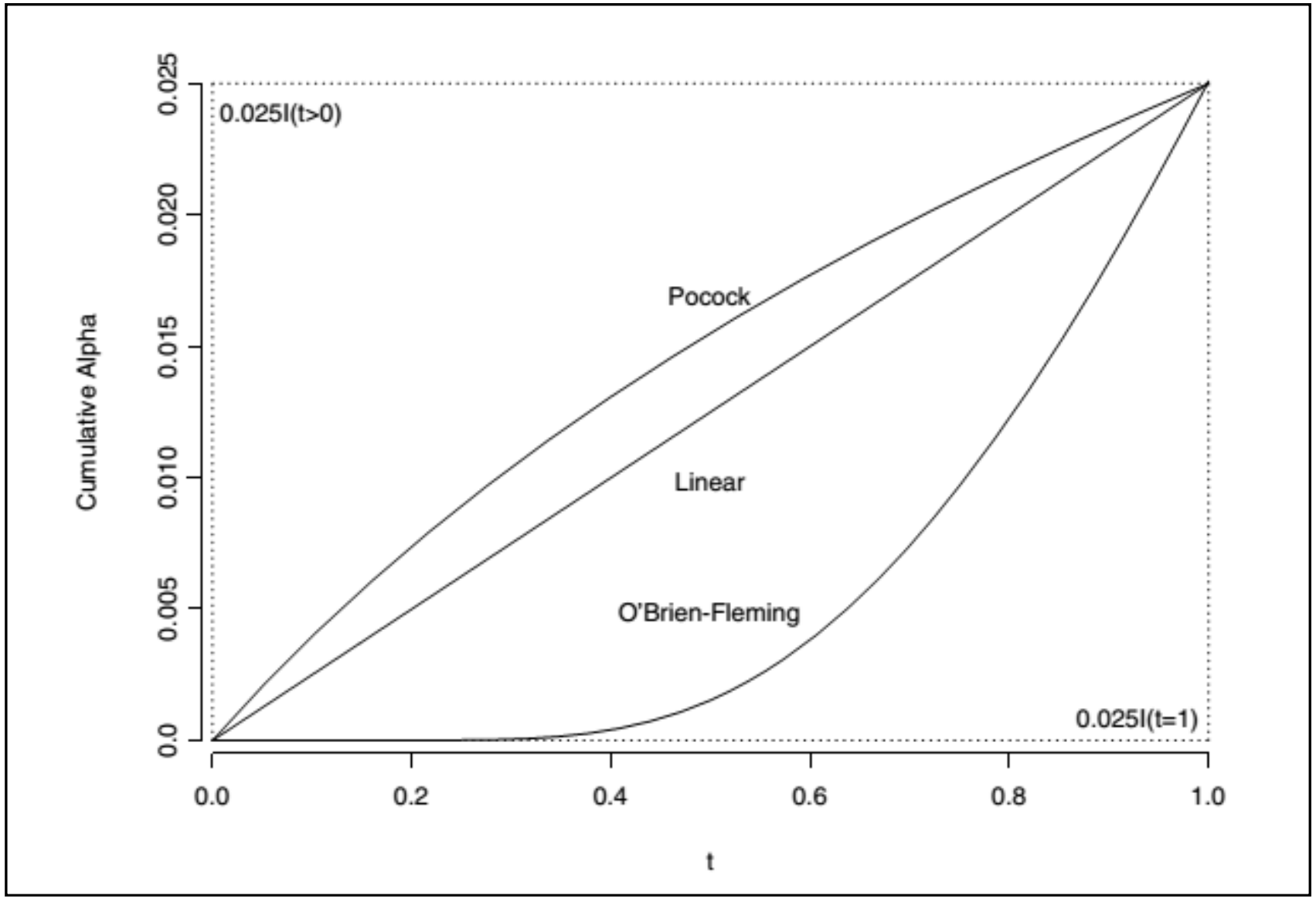
Wagenmakers, 2007


Optional stopping with 5 looks
Frequency of p-values
35000
30000
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
p-value

Sequential analysis controls Type 1 error rates (e.g., Pocock correction).

Because of the substantial savings in the expected number of observations effected by the sequential probability ratio test, and because of the simplicity of this test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy, at least for a certain period of time. The author was, therefore, requested to submit his findings in a restricted report [7] which was dated September, 1943.[3] In this

Wald, 1945


0.025I(t>0)
Pocock
Linear
O'Brien-Fleming
0.025I(t=1)
Cumulative Alpha
0.0
0.005
0.010
0.015
0.020
0.025
0.0
0.2
0.4
0.6
0.8
1.0
t

Pocock Boundary

Number of analyses	*p*-value threshold
2	0.0294
3	0.0221
4	0.0182
5	0.0158

Increase power by:

1) Decreasing measurement error

Increase power by:

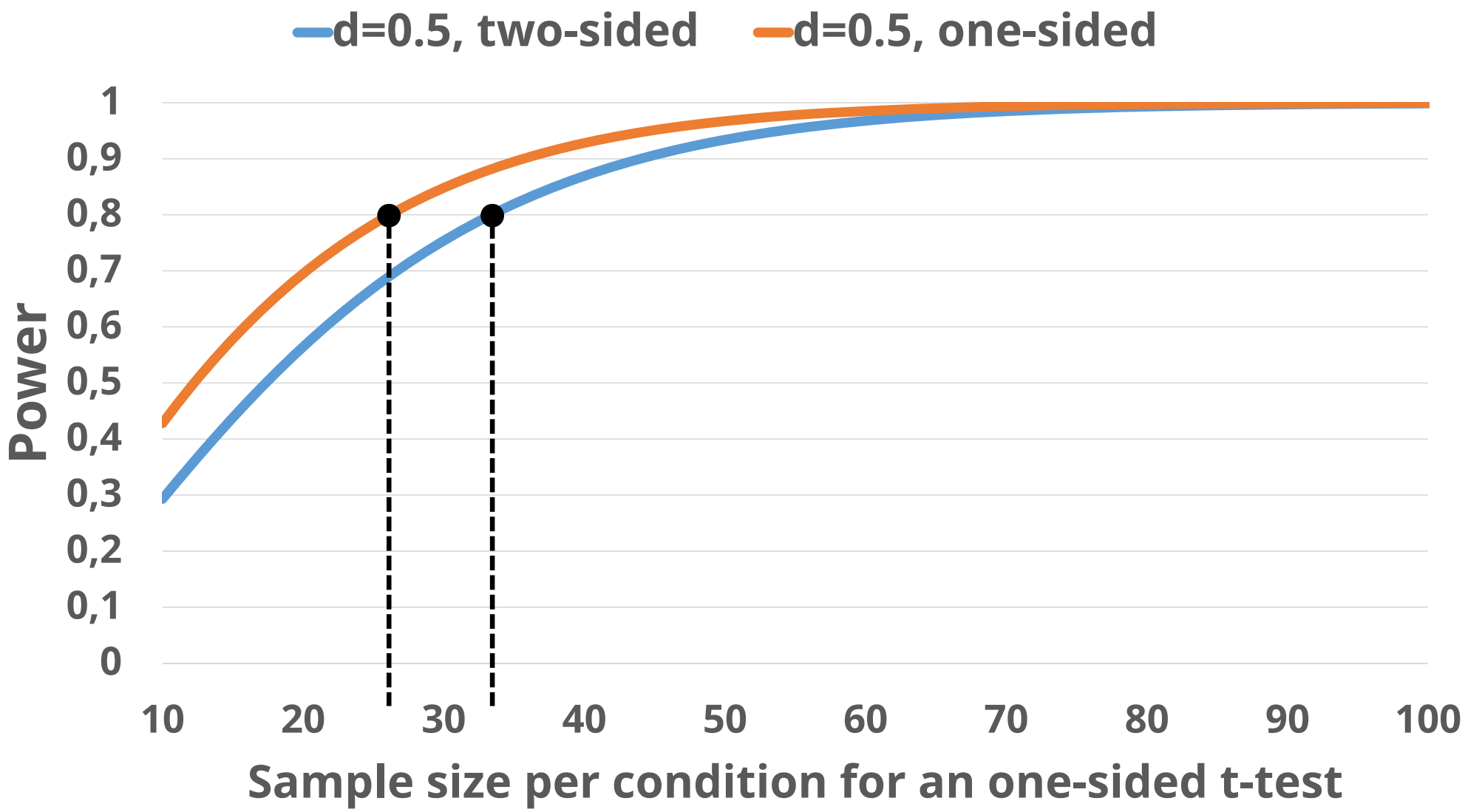
2) Using within designs (when within-measure correlation > 0.5)

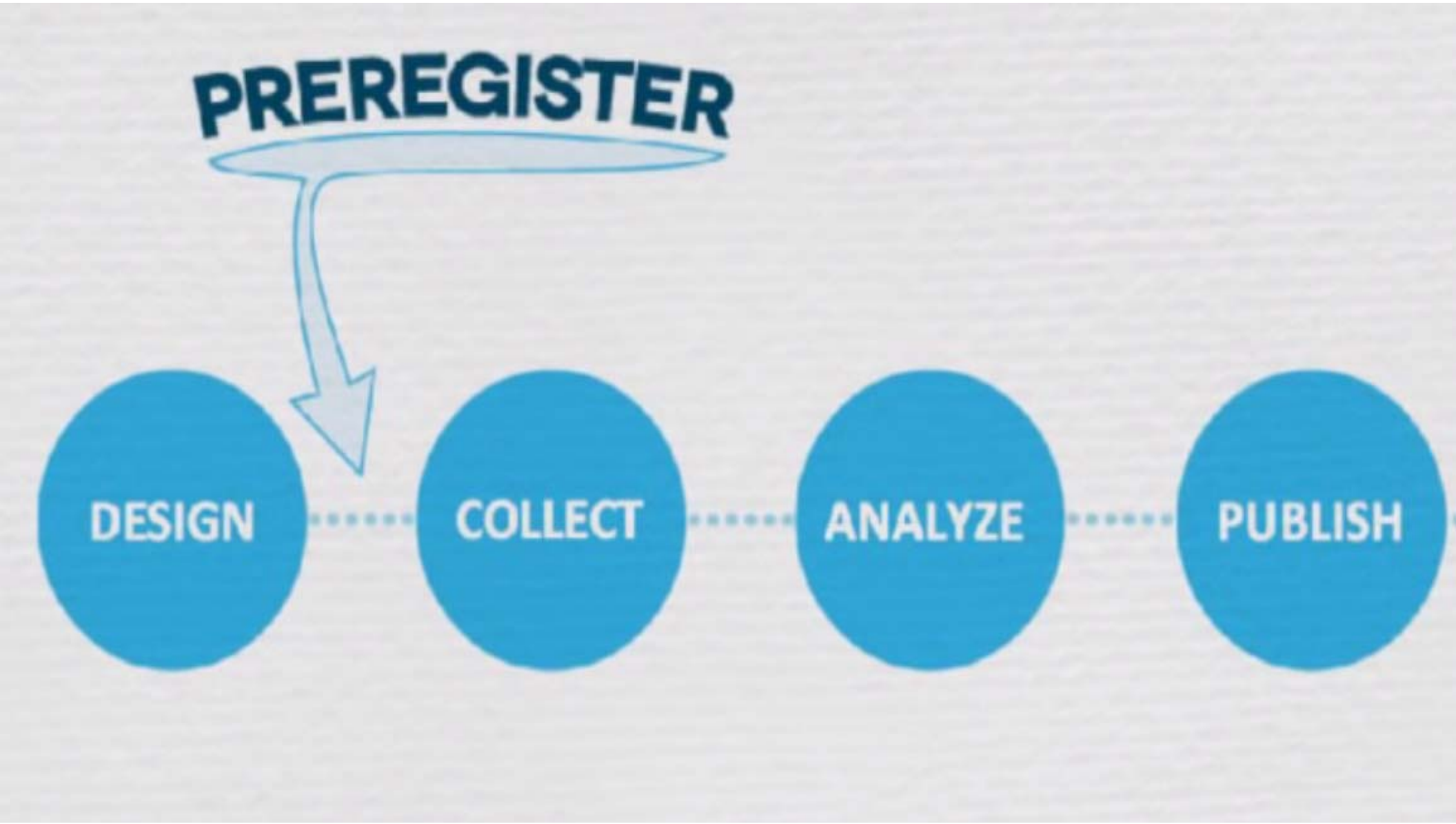
Increase power by:

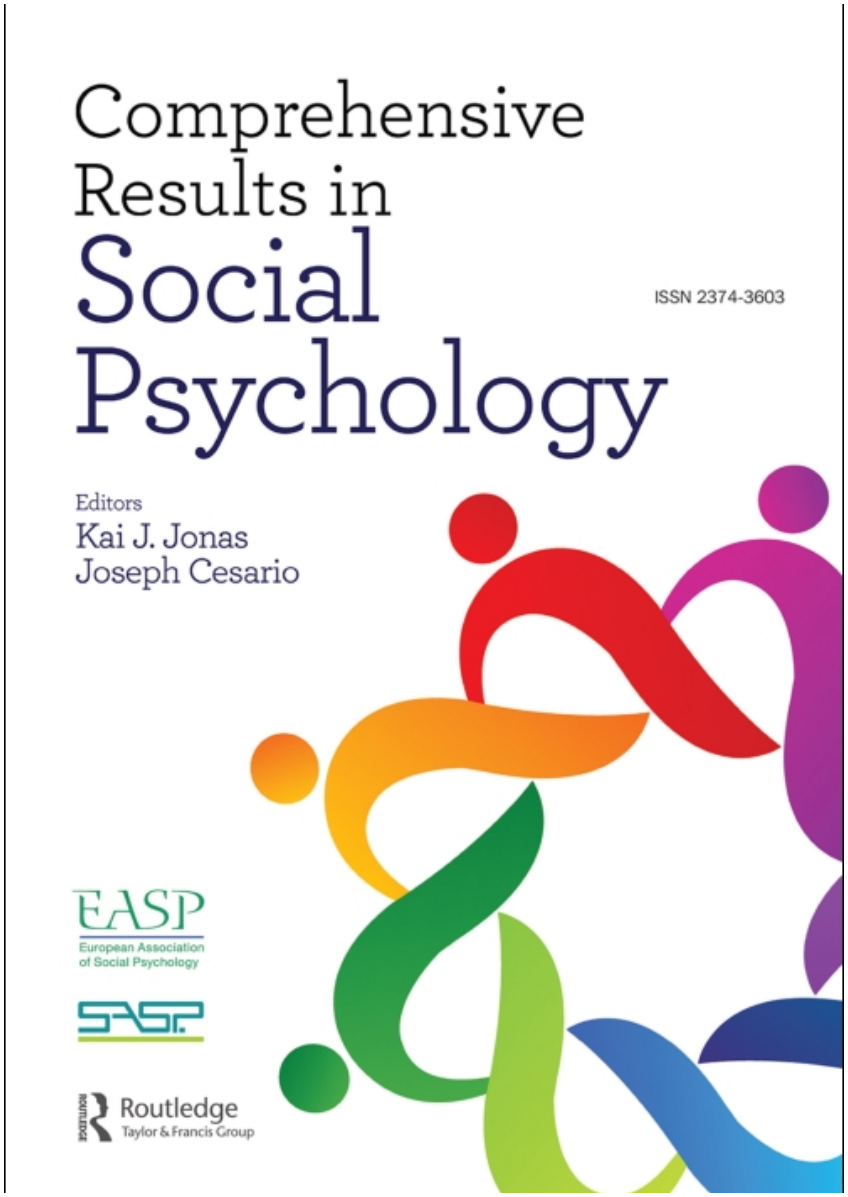
3) Increasing variability (e.g., use 7 or 9 instead of 3 or 5 answer scales)

Increase power by:

4) Use one-sided tests (if you have a directional prediction)


d=0.5, two-sided
d=0.5, one-sided
Power
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
10
20
30
40
50
60
70
80
90
100
Sample size per condition for an one-sided t-test

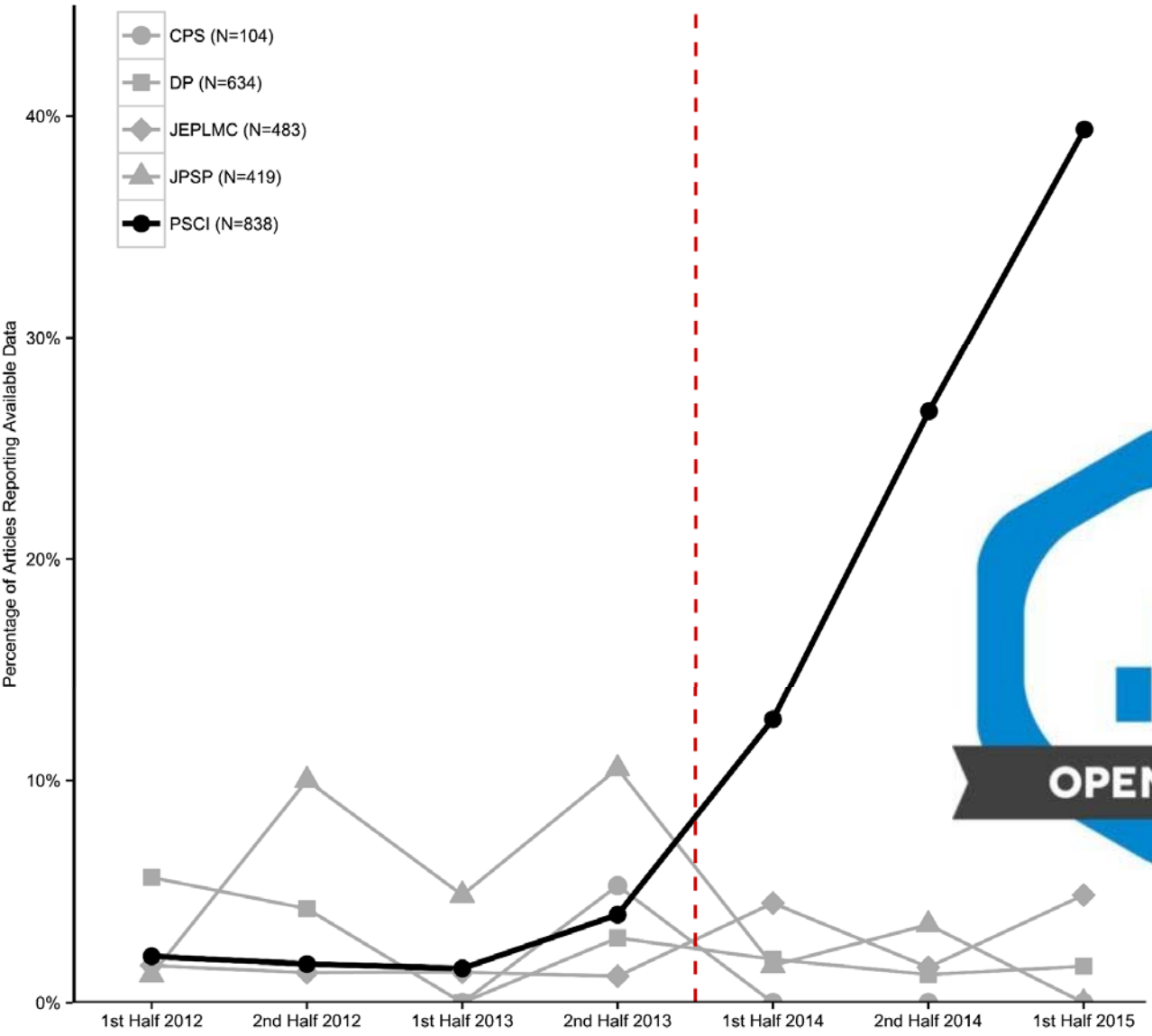

PREREGISTER
DESIGN
COLLECT
ANALYZE
PUBLISH


Comprehensive
Results in
Social
Psychology
ISSN 2374-3603
Editors
Kai J. Jonas
Joseph Cesario
EASP
European Association
of Social Psychology
SASP
Routledge
Taylor & Francis Group


ELSEVIER
JANUARY 2016 | VOLUME 74 | ISSN 0010-9452 |
SPECIAL SECTION:
WHAT'S YOUR POISON?
NEUROBEHAVIOURAL
CONSEQUENCES OF
EXPOSURE TO
INDUSTRIAL,
AGRICULTURAL AND
ENVIRONMENTAL
CHEMICALS
Cortex
A journal devoted to the study of the nervous system and behavior
Guest Editors:
Sarah Mackenzie Ross and
Virginia Harrison
Available online at www.sciencedirect.com
ScienceDirect

Share non-significant results (for novel and replication studies).

Share data, code and materials to make science more efficient


CPS (N=104)
DP (N=634)
JEPLMC (N=483)
JPSP (N=419)
PSCI (N=838)
Percentage of Articles Reporting Available Data
40%
30%
20%
10%
0%
1st Half 2012
2nd Half 2012
1st Half 2013
2nd Half 2013
1st Half 2014
2nd Half 2014
1st Half 2015


OPEN DATA


OPEN MATERIALS

SHERPA/RoMEO

Publisher copyright policies & self-archiving

Search

Journal titles or ISSNs Publisher names

Journal of Experimental Social Psychology

SHERPA/RoMEO

Search - Publisher copyright policies & self-archiving

One journal found when searched for: **journal of experimental social psychology**

Journal:	Journal of Experimental Social Psychology (ISSN: 0022-1031)
RoMEO:	This is a RoMEO **green** journal
Paid OA:	A paid open access option is **available** for this journal.
Author's Pre-print:	✓ author **can** archive pre-print (ie pre-refereeing)
Author's Post-print:	✓ author **can** archive post-print (ie final draft post-refereeing)
Publisher's Version/PDF:	✗ author **cannot** archive publisher's version/PDF
General Conditions:	

	Pro-Self (no sharing, file-drawer, p-hacking)	**Pro-Social (data sharing, replication, pre-registration)**
Pro-Self (no sharing, file-drawer, p-hacking)	+- / +-	- / +
Pro-Social (data sharing, replication, pre-registration)	+ / -	++ / ++

Thanks!

@Lakens

http://daniellakens.blogspot.nl/