



Conducting realistic, controlled experiments in software engineering

September, 2019
Magne Jørgensen
Simula Metropolitan

1

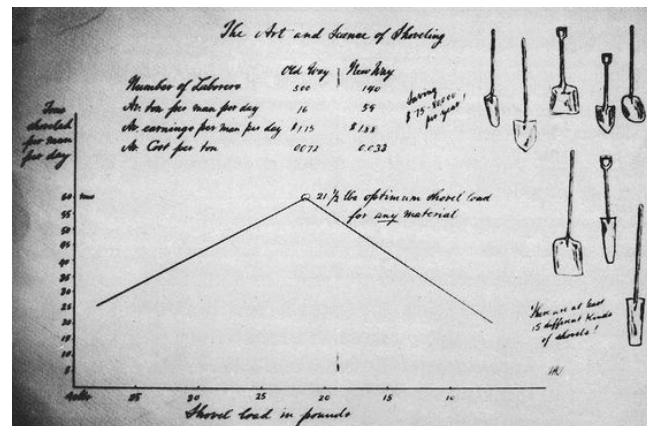
The beginning

- *Amicus Plato, sed magis amica veritas.* (Aristotle: "Plato is dear to me, but dearer still the truth")
- "History of animals" (350 B.C) by Aristotle may describe the first documented scientific experiment.
 - ✓ The development of the chicken from the egg
 - ✓ Experimental control: Daily opening of fertilized eggs to see the development of the chicken
- People have probably "experimented" with hunting techniques, food preparations, etc. long before that

2

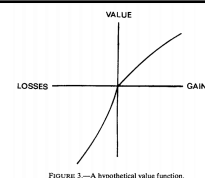
Famous experiments: Taylor

- Frederik Winslow Taylor (Scientific management, Taylorism, the first controlled experiments on work life)
 - ✓ Example: Experiments on the art and science of shoveling



3

Example from psychology: Prospect theory



- Study by Daniel Kahneman and Amos Tversky, 1979.
- Theory of **how people choose under uncertainty**, replacing (or competing with) the expected utility theory in economy
- Nobel prize in economy for this work in 2002
- Very simple and artificial experiments with psychology students, e.g., What would you choose:
 - ✓ A: 50% chance to win 1.000 and 50% chance to win nothing, OR
 - ✓ B: 450 for sure
- No real-world experiments on how people actually choose. Hypothetical questions. No sample-to-population type of generalization.
- We would classify this as having very low external validity, when thinking in terms of sample-to-population generalization.
- Extreme focus on the core mechanisms. Many experiments in one study to understand the mechanisms.
- ... and, some field studies later on to test the theory ;-)

4

We should examine core mechanism more often

- Less focus on latest fashion
- Fashion focus typically means that we'll be too late to have a strong impact
 - ✓ E.g., research on agile methods after agile has been implemented everywhere is hard to affect.
- Fashion focus also means that the results soon will be outdated
 - ✓ E.g., research done on RUP. Not of much value today.
- We should generalize by understanding core mechanisms, not by sample-to-population
 - ✓ External validity discussion will then also change and become more meaningful
- Should not be afraid of studies in artificial contexts, as long as they increase understanding of the core mechanisms.

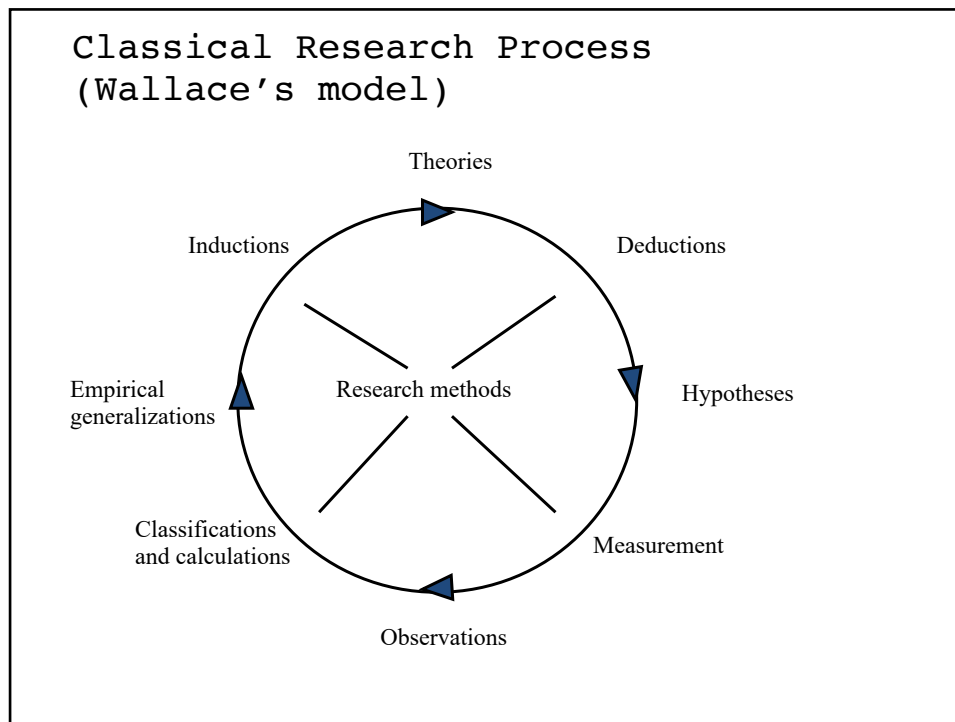
5

Famous experiments: Fisher introduces p-values and statistical hypothesis testing

- Fisher (Randomizing, p-values, "A lady tasting tea")
 - ✓ A lady claims that she can taste whether the milk or the was added first
 - ✓ How many cups are needed to test the claim?
 - ✓ How would you design the experiment?
 - ✓ Is there any problems with the experiment?



6



7

Are controlled experiments better than case studies, surveys, interpretative research, ...?

- Wrong question! Most research methods and paradigms have their strengths and weaknesses.
- It is the fit between the research method and research question (goal of a study) that matters.
- In practice, unfortunately, the choice of research method and paradigm is very much determined by personal preference and/or set of personal values (ideals).
 - ✓ This has the consequence that the choice of research method may be value based instead of selection of the best suited research methods.
 - ✓ Researchers belonging to "interpretive research" tend to be overly negative about the use of experiments and statistics on people.
 - ✓ Researchers belonging to "positivism" tend to be overly negative regarding the value of case studies, lack of "control group" and no prior hypotheses.

8

What is an experiment?

- Manipulation of at least one variable, usually termed the "treatment".
- Usually, but not necessarily, testing hypotheses, beliefs, theories etc. stated upfront.
 - ✓ Often (much too often!) testing whether there is an effect from X (rejecting a "null hypothesis"), but could also (and should more often) try to answer questions like "How large is the effect of X?"
- Method strong on finding cause-effect relationships, especially when treatment is randomized.
- Without randomized treatment we have a *quasi-experiment* where we have to use other argumentation than the experiment itself to convince ourselves and others that there are no alternative explanations than the treatment to explain the effect.
 - ✓ **Example:** A company let their managers select between the use of "fixed price" and "hourly paid" (contract types) projects. The hourly paid projects performed much better. Cause-effect or correlation? (My current – ongoing – experiment on this topic)

9

Typical experimental process

- Description of a problem/challenge/theory to be tested
- Hypothesis relevant for addressing the problem/challenge/theory
 - ✓ For example: Treatment A leads to higher performance than treatment B.
- Design of a study where the hypothesis can be tested.
 - ✓ Study may be designed to analyse the existence of an effect of treatment, examine typical effect size or something else.
 - ✓ Study context (task, people, tools) may be representative, extreme, randomly selected, or convenience samples.
- Allocation of treatment to subjects
 - ✓ Randomly, self-selected, ("natural" selection), ...
- Execution of study, measurement and data collection
- Statistical analysis of data.
 - ✓ For example: Is the difference in effect statistically significant? What is the 90% confidence interval of the effect size?
- Interpretation of results - in light of the collected and analyzed data, understanding of mechanisms and previous results. Not the data alone!

10

Important quality principles of experiments

- Controlled comparison (control group, baseline)
 - ✓ Avoiding to claim treatments effects that would happen anyway (Placebo effects, natural learning etc.)
- Randomization (not to be confused with representativeness of the sample or haphazardly allocation)
 - ✓ Randomization is about internal validity
 - ✓ Representativeness is about external validity
- Blinded analysis, blinded treatment (typically difficult and seldom used in software engineering experiments)
- Replications/reproducible results. Repeated measurement, or at least enabled to be replicated by others.
- Simplicity of design. **Optimally only one question per experiment.** As simple design as possible. Should have very good reasons to use a complex experiment design.

11

Internal validity of experiments

- Does the experiment warrant that the treatment causes the effect (Are there alternative explanations that can explain the results?)
 - ✓ Events other than the treatment that could have impacted the outcome?
 - ✓ Fatigue confounded the effect of the treatment?
 - ✓ Hawthorne effect? (which really should have another name)
 - ✓ Measurement problems?
 - ✓ Statistical regression?
 - ✓ Biased selection of subjects, or biased allocation of subjects to treatment?
 - ✓ Different loss of participants in different treatment groups?

12

External validity of experiments

- Are the results representative for the population of interest (the one we want to generalize to)?
 - ✓ Participants (When are students sufficiently representative of software professionals?)
 - ✓ Tasks (What can we say about real world tasks based on results from smaller tasks?)
 - ✓ Contexts (What can we say about real-world effects from effects in laboratory settings?)
- Sampling process essential. (Allocation of treatment-process is also important, since external validity requires internal validity).
 - ✓ Induction from few to many (generalizing) in software engineering nearly always require additional analytic argumentation.
- Not satisfactory to say that "*results only applies to what we have studied*". An argumentation/analysis of generalization is always (?) needed.

13

Be aware of your own (and others)
researcher and publication bias

14

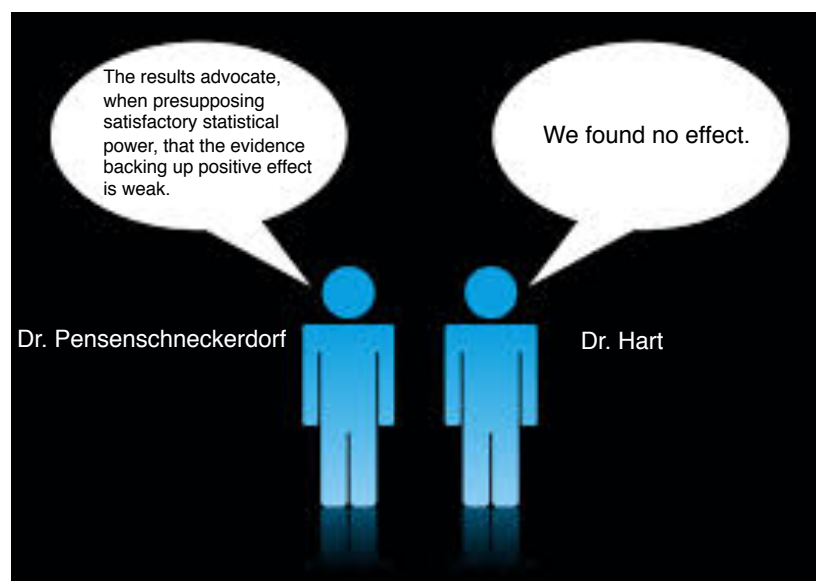
Illustration

How easy is it to find statistically significant results With “flexible” research practise?

I made a simple experiment ...

15

My hypothesis: Researchers with longer names write more complex papers



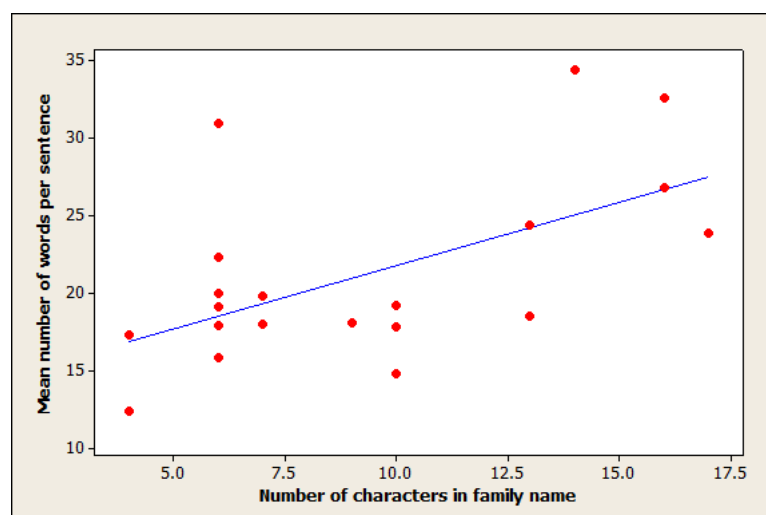
16

Design and results of study

- **Variables:**
 - ✓ LengthOfName: Length of surname of the first author
 - ✓ Complexity1: Number of words per paragraph
 - ✓ Complexity2: Flesch-Kincaid reading level
- **Data collection:**
 - ✓ The first 20 publications identified by “google scholar” using the search string “software engineering” for year 2012. (n=20 is a typical sample size for software engineering studies)
- **Results were statistically significant!**
 - ✓ $r_{\text{LengthOfName,Complexity1}} = 0.581$ ($p=0.007$)
 - ✓ $r_{\text{LengthOfName,Complexity2}} = 0.577$ ($p=0.008$)
- **Conclusion:** The analysis reject the null-hypothesis that there is no difference, i.e., a support of that long names are connected with more complex texts. Results can be published!?

17

The regression line also supports that there is a relation between length of name and complexity of writing



18

How did I do it?

(How to easily get “interesting” results in any study)

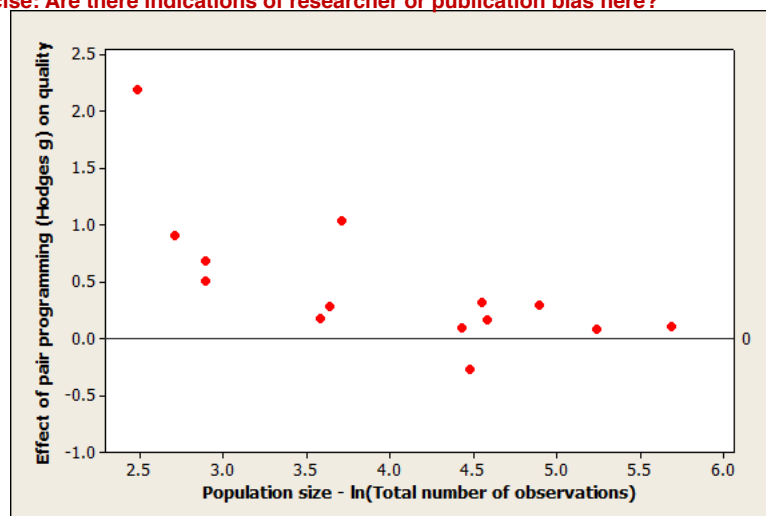
- **Publication bias:** Only two, out of fourteen, significant measures of paper complexity were reported.
- **Researcher bias 1:** A (defendable?), post hoc (after looking at the data) change in how to measure name length.
 - ✓ The use of surname length was motivated by the observation that not all authors informed about their first name.
- **Researcher bias 2:** A (defendable?), post hoc removal of two observations.
 - ✓ Motivated by the lack of data for the Flesh-Kincaid measure of those two papers.
- **Low number of observations:** Statistical power approx. 0.3 (assuming $r=0.3$, $p<0.05$).
 - ✓ If research was a game where the winners have $p<0.05$, 5 studies with 20 observations is much better than one with 100.

19

Effect sizes in studies on pair programming

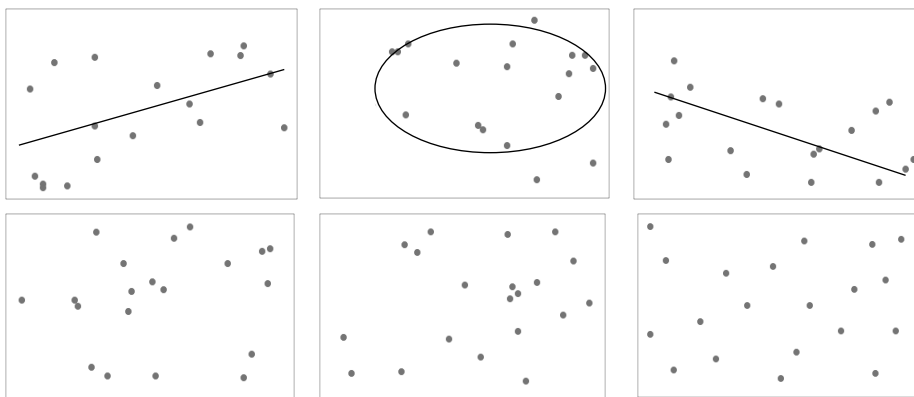
Source: Hannay, Jo E., et al. "The effectiveness of pair programming: A meta-analysis." Information and Software Technology 51.7 (2009): 1110-1122.

Exercise: Are there indications of researcher or publication bias here?



20

Finding relationships in randomness ...



How many would show a pattern if allowed to remove 1-2 "outliers"?
 (Only the last one is non-random. The first five are the first five
 I got from a random data generator.)

21

When are experiments more likely to give incorrect results?

- Low sample size (low statistical power)
- Small (true) effect size (low statistical power, unless very large sample size)
- High number of relationships tested, and selective reporting (publication bias)
- High flexibility in design and interpretations, e.g., flexibility related to measures, statistical tests, study design, model tuning, definition of outliers, interpretation of data (researcher bias)
- Substantial degree of vested interests or wish for a particular outcome (researcher bias)
- Hot scientific topic (researcher bias).

22

All experiments should start with an answerable question

- What is "wrong" with the following questions?
 - ✓ How much heat is lost through the head?
 - Claim: Most heat is lost through the head
 - ✓ How much of academic skill is "nature" and how much is "nurture"?
 - Claim: Academic skill is 60% nature (inheritance) and 40% nurture (environment).
 - ✓ What is the difference in productivity between programmers?
 - Claim: 1:10
 - ✓ Is agile development better than traditional methods?
 - Claim: Agile is better than traditional methods

23

What is a meaningful question?

- A few criteria:
 - ✓ Answer a relevant question (not whether there is a difference between A and B with respect to X, but rather how large the difference is)
 - ✓ Sufficiently precise (defines, for example, the meaning of "better than")
 - ✓ Sufficiently context-dependent ("when is A better than B, rather than "is A better than B")
 - ✓ Not too complex phenomenon (the effect of an approach composed of several methods in a variety of contexts may be impossible to analyse properly)

24

Common experimental designs

- One group pretest-posttest design
 - ✓ Performance 1, Treatment, Performance 2
- Completely randomized design:
 - ✓ Random allocation of treatment to subjects
- Randomized block design:
 - ✓ Similar to completely randomized design, but considering block's effect e.g., using "Latin-square"
 - ✓ Useful when completely randomized risk different sample size in each group
- Crossover design: Matched (paired) analysis of variance
 - ✓ Random allocation to Group A or B
 - ✓ Group A: Treatment 1, Performance 1, Treatment 2, Performance 2
 - ✓ Group B: Treatment 2, Performance 1, Treatment 1, Performance 2
- **NB:** The statistical analysis should fit the type of design

25

Inferential (inductive) statistical methods

- Null Hypothesis Testing
 - ✓ A hybrid of conflicting methods
 - ✓ Complex to interpret. Not really answering your questions.
- Confidence intervals
 - ✓ Gives more useful information and answer more meaningful questions
 - ✓ Complex to interpret. Not really answering your questions, either.
- What to do?
 - ✓ Use confidence intervals proper for your experimental design (and possible p-values), but be aware of their limitations.
 - ✓ Visualize the effects using the actual data.
 - ✓ Include (if possible) meta-analyses with other data, or at least discuss your results in light of other data.

26

The end

27

EXTRA MATERIAL

28

Positivist research

- Originally developed for the use in natural science, i.e., not studies of human behavior.
- Knowledge generation through Wallace's cycle (see next slide).
- Based on reductionism, repeatability and refutation (falsification, ref. Popper).
- Assumptions:
 - ✓ Our world is ordered, not random
 - ✓ We can investigate the world objectively (Well, at least achieve an acceptable degree of "inter-subjectivity".)
- Goal: Discover patterns.
- Criteria:
 - ✓ Objectivity (or at least inter-subjectivity)
 - ✓ Reliability
 - ✓ Internal validity (= the extent to which a study evaluates the intended hypotheses, i.e., that it is not likely that rival hypotheses explains the findings)
 - ✓ External validity (= the extent to which the results of a study extend beyond the limited sample used in the study)

29

Interpretivist research

- *"Interpretive research in IS and computing is concerned with understanding the social context of an information system: the social processes by which it is developed and construed by people and through which it influences, and is influenced by, its social setting."* (p 292, in Briony J. Oates)
- Try to identify, explore and explain ("rich understanding") how factors in a particular social setting are related and interdependent. Case studies are typically preferred.
- Characteristics:
 - ✓ Multiple subjective realities
 - ✓ Dynamic, socially constructed meaning
 - ✓ Researcher reflexivity (researchers should reflect on their own assumptions, beliefs and actions and their impact on the research process)
 - ✓ Study of people in their natural social setting (typically, case studies)
 - ✓ Qualitative data analysis
 - ✓ Multiple interpretations

30

Interpretive Research

- Criteria (somewhat forced into a positivistic framework):
 - ✓ Trustworthiness (more general concept than validity?)
 - ✓ Confirmability (analogue to objectivity – can we follow the arguments from the raw data to the interpretation?)
 - ✓ Dependability (analogue to reliability and repeatability – is the research process well documented?)
 - ✓ Credibility (analogue to internal validity – is it valid to draw the conclusions based on the data collected?)
 - ✓ Transferability (analogue to external validity – is it possible to transfer the findings to other cases?
 - **NB:** This is frequently not a goal in interpretive research. An interesting case is an interesting case, even when not transferable to other cases.

31

Quantitative vs qualitative research

- Possible differences:
 - ✓ Statistical analysis vs interviews and text analyses?
 - ✓ Experiments vs case studies?
 - ✓ Positivistic vs interpretivistic?
 - ✓ "Natural science" vs "Social science"?
- The subject of study should decide the selection of research method, e.g.
 - ✓ a study of the connection between education and salary levels may be hard to carry out without measurements.
 - ✓ a study of how people perceive the power structure at universities may be hard to carry out without talking to people and analysing their answers and/or observing behavior.
- **Remember:** There are no good or bad research methods, only good and bad research and different degrees of fit between research method and research problem.

32

Main types of qualitative research methods

- Observation
 - ✓ Example: Observation of people's behavior at meetings
- Analysing texts and documents
 - ✓ Example: Analysing project experience reports and minutes from status meetings
- Interviews
 - ✓ Example: Interviews with experienced managers about the reasons of estimation errors.
- Recording and transcribing
 - ✓ Example: Videorecording of team work. Categorizing the communication according to types of statements.

33

Are most published research findings in empirical software engineering wrong or with exaggerated effect sizes?
How to improve?

Magne Jørgensen
ISERN-workshop
20 October, 2015



34

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Most Research Findings Are False for Most Research Designs and for Most Fields

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework a research finding is less likely to be true when the studies conducted in a field are smaller when effect sizes are smaller when there is a greater number and lesser selection of tested relationships, where there is greater flexibility in designs, definitions, outcomes, and analytical modes when there is greater financial and other interest and prejudice and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *positive*. Research findings need defined

is characterized by a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us first consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the powers similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that r relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true

probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R / (R - \beta R + \alpha)$. A research finding is three

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

35



OCTOBER 19TH - 27TH 2013 Economist.com

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

99
Einsteinium

36

Nature, October 2015, Regina Nuzzo



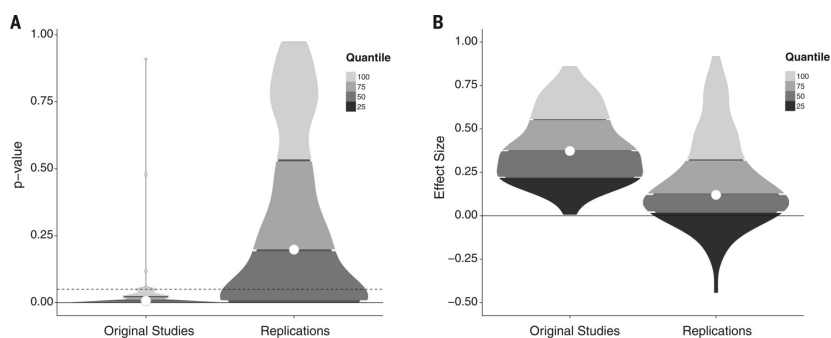
FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION.
 BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY
 RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

37

PSYCHOLOGY: Independent replications, with high statistical power, of 100 randomly selected studies gave shocking results!

Reference: Open Science Collaboration. Estimating the reproducibility of psychological science. Science 349.6251 (2015): aac4716.



If we did a similar replication exercise in empirical software engineering (maybe we should!), what would we find?

38

Our study indicates that we will find similarly disappointing results in empirical software engineering

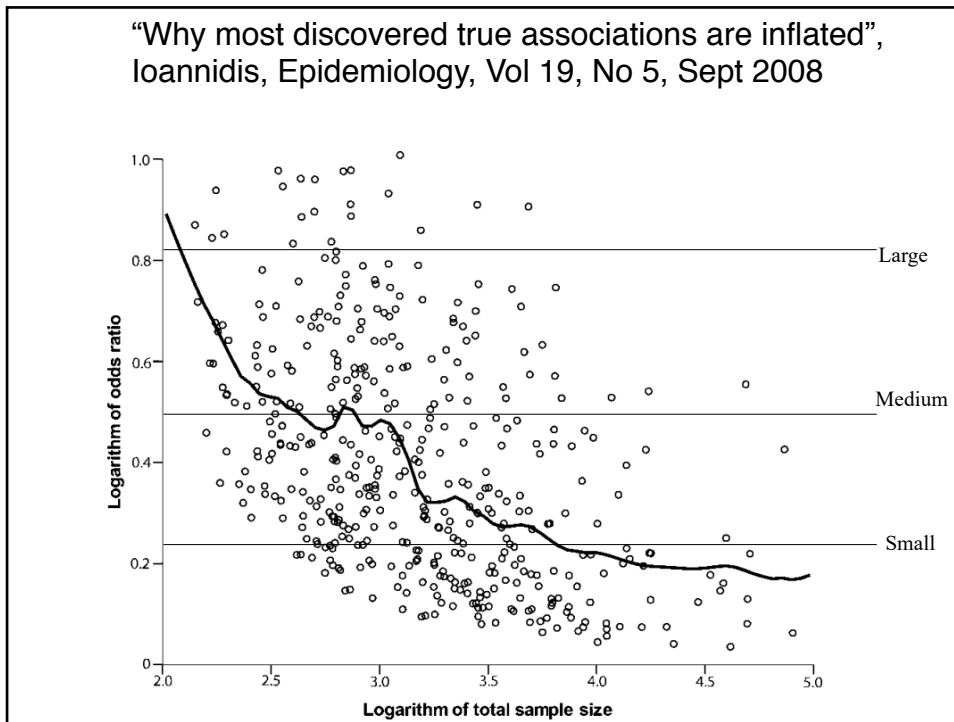
Based on calculations of amount of researcher and publication bias needed to explain the high proportion of statistically significant results given the low statistical power of SE studies.

Jørgensen, M., Dybå, T., Liestøl, K., & Sjøberg, D. I. (2015). Incorrect results in software engineering experiments: How to improve research practices. To appear in Journal of Systems and Software.

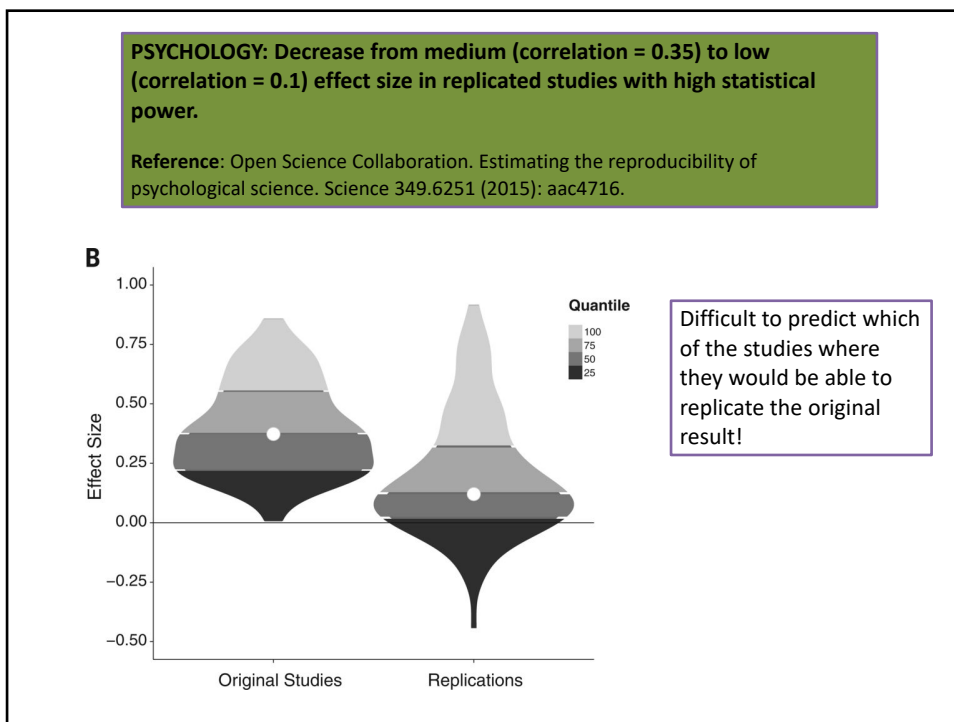
39

EXAGGERATED EFFECT SIZES OF SMALL STUDIES

40



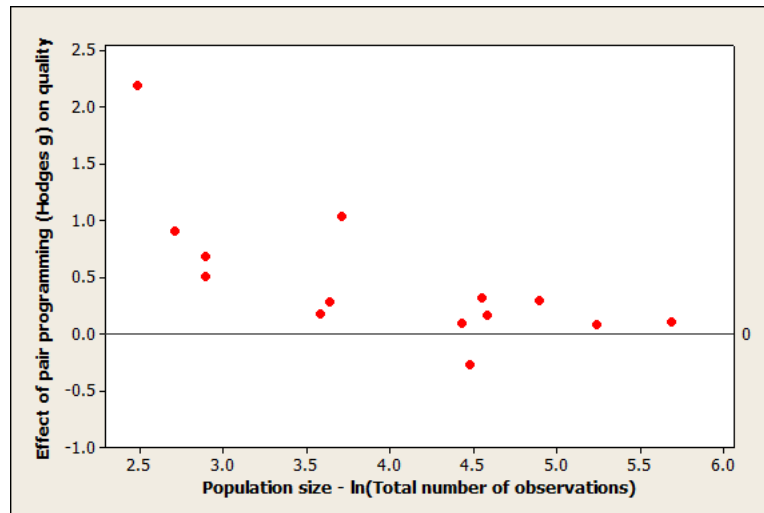
41



42

Example from software engineering: Effect sizes from studies on pair programming

Source: Hannay, Jo E., et al. "The effectiveness of pair programming: A meta-analysis." Information and Software Technology 51.7 (2009): 1110-1122.



43

The typical effect size in empirical SE studies

- Previously reported median effect size of SE experiments suggests that it is medium ($r=0.3$), but did not adjust for inflated effect size.
 - ✓ Kampenes, Vigdis By, et al. "A systematic review of effect size in software engineering experiments." Information and Software Technology 49.11 (2007): 1073-1086.
- **Probably the true effect sizes in SE are even lower than previously reported, e.g., between small and medium (r between 0.1 and 0.2).**

44

LOW EFFECT SIZES
 + LOW NUMBER OF SUBJECTS
 = VERY LOW STATISTICAL POWER

45

Average power of SE studies of about
 0.2?
 (best case of 0.3)

Frequency and cumulative percentage distribution of power in 92 controlled SE experiments

Power level	Small effect size		Medium effect size		Large effect size	
	Freq.	Cum. %	Freq.	Cum. %	Freq.	Cum. %
.91-.99	–	–	18	100	69	100
.81-.90	1	100	11	96	75	85
.71-.80	–	100	14	94	49	69
.61-.70	2	100	13	91	70	58
.51-.60	9	99	44	88	58	43
.41-.50	2	97	50	78	21	30
.31-.40	–	97	76	67	43	25
.21-.30	13	97	107	51	43	16
.11-.20	120	94	94	27	31	7
.00-.10	312	68	32	7	–	–
Total	459	–	459	–	459	–
Average power	0.11	–	0.36	–	0.63	–

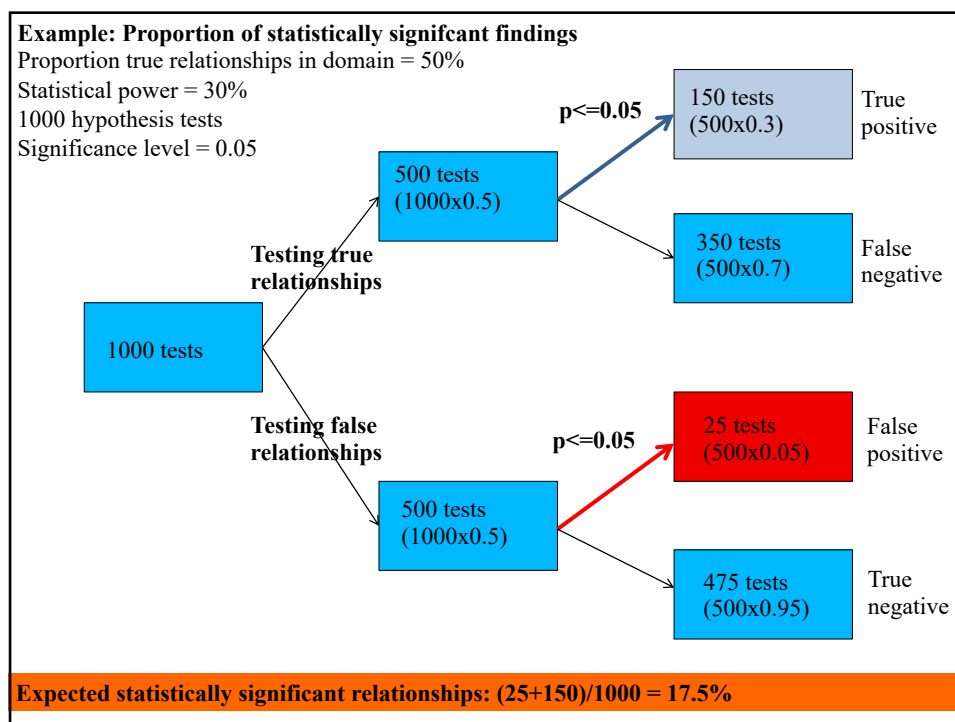
Dybå, Tore, Vigdis By Kampenes, and Dag IK Sjøberg. "A systematic review of statistical power in software engineering experiments." *Information and Software Technology* 48.8 (2006): 745-755.

46

20-30% statistical power means that
 With 1000 tests on real differences,
 only 2-300 should be statistically
 significant.

... in reality many of the tests will not
 be on real differences and we should
 expect much fewer than 2-300
 statistically significant results.

47



48

WHAT DO YOU THINK THE
ACTUAL PROPORTION OF
 $P < 0.05$ IN SE-STUDIES IS?

49

Proportion statistical significant results
Theoretical: Less than 30% (around 20%)
Actual: More than 50%!

Table 6: Results from the review

	Total	2002– 2003	2004– 2005	2006– 2007	2008– 2009	2010– 2011	2012– 2013
No. papers	150	25	25	25	25	25	25
No. experiments	196	30	31	32	37	35	31
Median sample size	29	47	33	32	23	26	27
No. hypothesis tests	1279	212	210	251	220	215	171
$p < 0.05^1$	52%	53%	59%	52%	46%	52%	54%
$p < 0.01^2$	27%	25%	30%	30%	25%	28%	23%

50

HOW MUCH RESEARCH AND PUBLICATION BIAS DO WE HAVE TO HAVE TO EXPLAIN A DIFFERENCE BETWEEN 20% EXPECTED AND 50% ACTUALLY OBSERVED STATISTICALLY SIGNIFICANT RELATIONSHIPS?

AND HOW DOES THIS AFFECT RESULT RELIABILITY?

51

Example of combinations of research and publication that lead to about 50% statistically significant results in a situation with 30% statistical power (the optimistic scenario)

Table 8: Expected median proportions of significant findings

		Researcher bias (rb)					
		0	0.1	0.2	0.3	0.4	0.5
Publication bias (pb)	0	23%	30%	38%	46%	54%	61%
	0.1	24%	33%	41%	48%	56%	63%
	0.2	27%	35%	43%	51%	59%	66%
	0.3	29%	38%	47%	55%	62%	69%
	0.4	33%	42%	50%	58%	66%	72%
	0.5	37%	46%	55%	63%	70%	76%
	0.6	42%	52%	60%	68%	74%	80%
	0.7	49%	59%	67%	74%	79%	84%
	0.8	59%	68%	75%	81%	85%	89%

52

The effect on result reliability ...

Domain with	Incorrect results (total)	Incorrect significant results
50% true relationships	Ca. 40%	Ca. 35%
30% true relationships	Ca. 60% (most results are false!)	Ca. 45% (nearly half of the significant results are false)

Indicates how much the proportion of incorrect results depends on the proportion true results in a topic/domain.

Topics where we test without any prior theory or good reason to expect a relationship consequently gives much less reliable results.

53

Practices leading to research and publication bias

TABLE 1: Results from a survey on statistical practices

Research Practice	Have experienced/done this in my own research				
	Never	Seldom	Occasionally	Often	Don't know
P1: Paper rejected due to non-significance ¹	14	6	8	4	4
P2: Paper not submitted due to non-significance ²	16	6	8	4	1
P3: Not reported non-significant results ³	17	8	4	4	2
P4: Not reported undesired results ⁴	18	8	0	4	4
R1: Post hoc hypotheses ⁵	11	4	12	6	1
R2: Post hoc outlier criteria ⁶	14	5	9	3	3
R3: Flexible reporting of measures and analyses ⁷	10	10	5	7	2

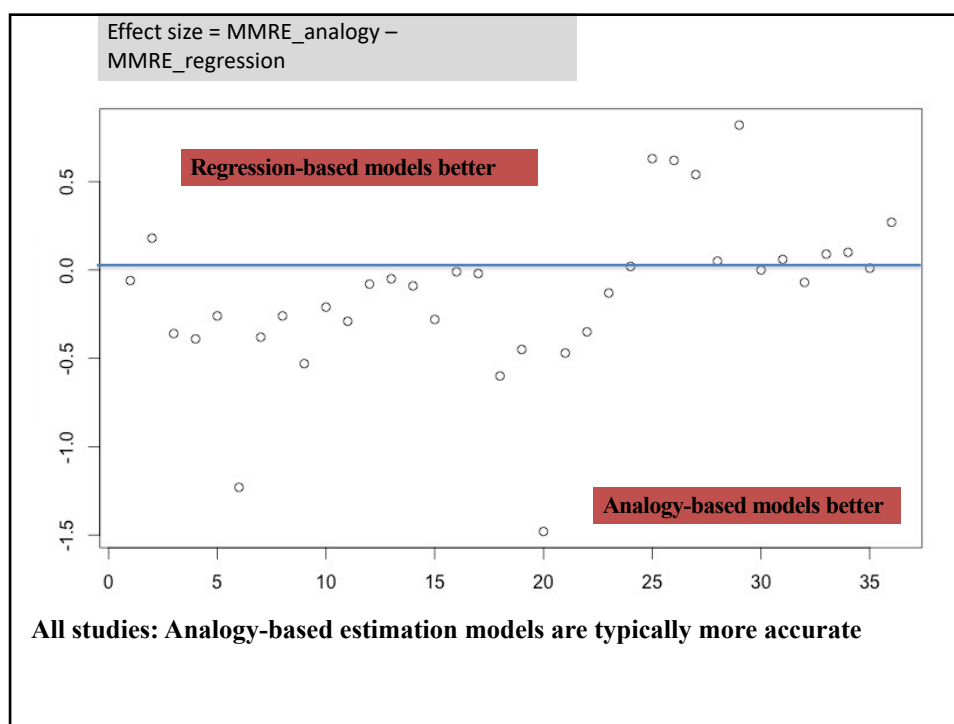
- “It's extremely hard to publish a journal paper without 'massaging' the data and the hypotheses first. If you do not do this, you will end up with no publications at all. I think journal editors and reviewers should do something, so that they encourage honest accounts of empirical work, and make researchers with non-significant results feel welcome.”
- “... unless authors do something really stupid, it's very easy to get away with post-hoc interventions. Sneaking up and making it to a journal publication is common and if many fellows practice it, why should we discriminate against ourselves by discarding the practice? The price appears to be too high for this.”

54

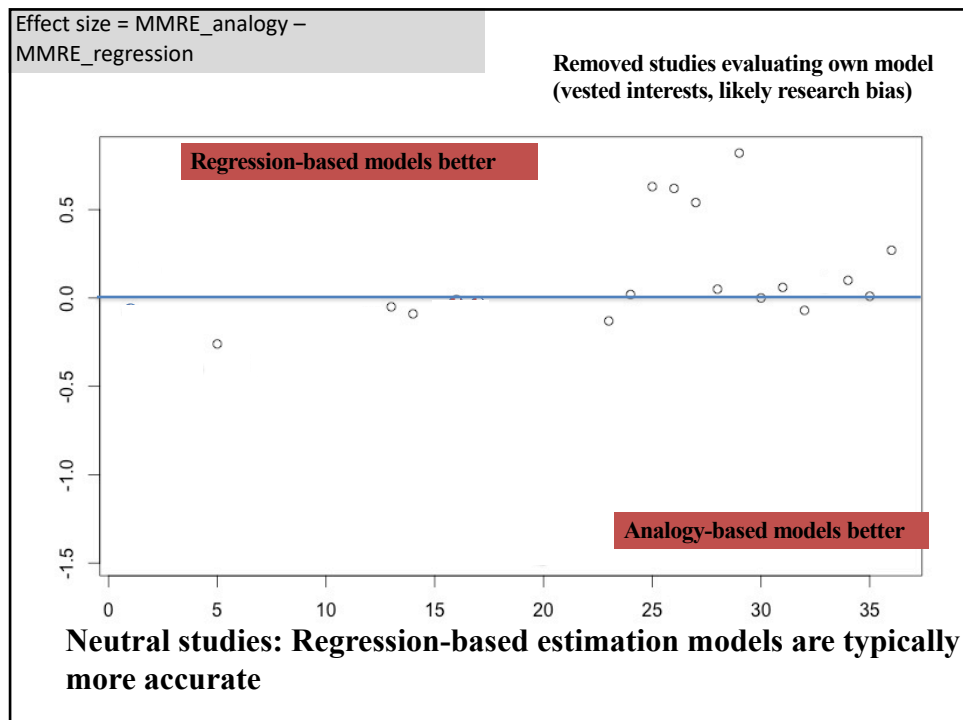
HOW MUCH RESEARCHER BIAS IS THERE?

EXAMPLE: STUDIES ON REGRESSION VS ANALOGY- BASED COST ESTIMATION MODELS

55



56



57

State-of-practice summarized





- Unsatisfactory low statistical power of most software engineering studies
- Exaggerated effect sizes
- Substantial levels of questionable practices (research and/or publication bias)
- Reasons to believe that at least (best case) one third of the statistically significant results are incorrect
 - ✓ Difficult to determine which result that are reproducible and which not.
- We need less "shotgun" type of hypothesis testing and more hypotheses based on theory and prior explorations ("less is more" when it comes to hypothesis testing)

58





HOW SCIENTISTS FOOL THEMSELVES — AND HOW THEY CAN STOP

Humans are remarkably good at self-deception. But growing concern about reproducibility is driving many researchers to seek ways to fight their own worst instincts.

COGNITIVE FALLACIES IN RESEARCH

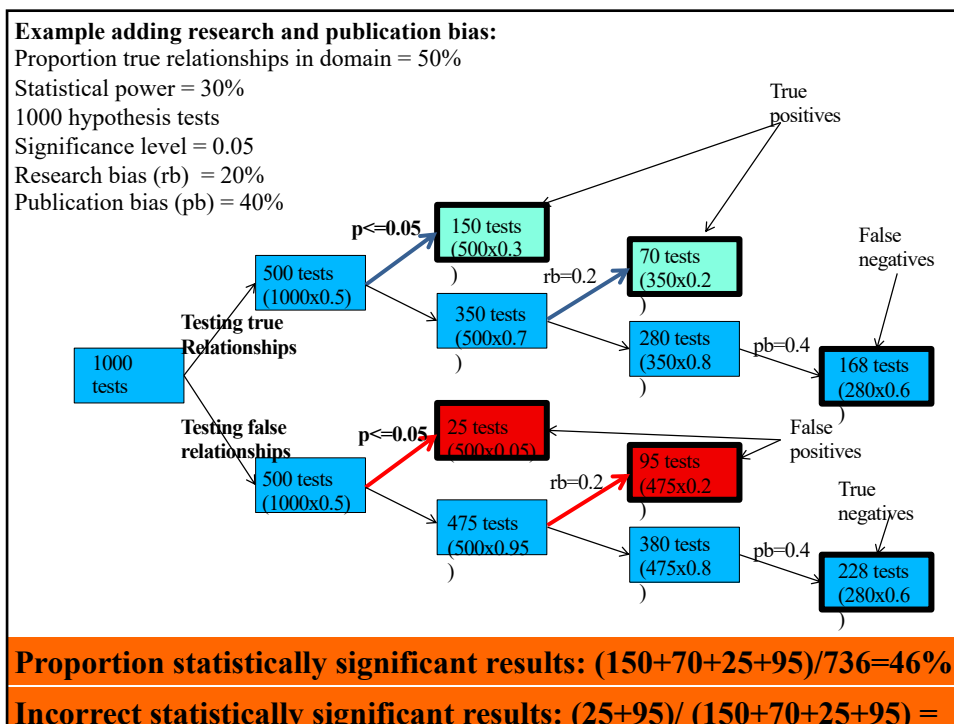
 HYPOTHESIS MYOPIA Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.	 TEXAS SHARPSHOOTER Seizing on random patterns in the data and mistaking them for interesting findings.	 ASYMMETRIC ATTENTION Rigorously checking unexpected results, but giving expected ones a free pass.	 JUST-SO STORYTELLING Finding stories after the fact to rationalize whatever the results turn out to be.
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

DEBIASING TECHNIQUES

 DEVIL'S ADVOCACY Explicitly consider alternative hypotheses — then test them out head-to-head.	 PRE-COMMITMENT Publicly declare a data collection and analysis plan before starting the study.	 TEAM OF RIVALS Invite your academic adversaries to collaborate with you on a study.	 BLIND DATA ANALYSIS Analyse data that look real but are not exactly what you collected — and then lift the blind.
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

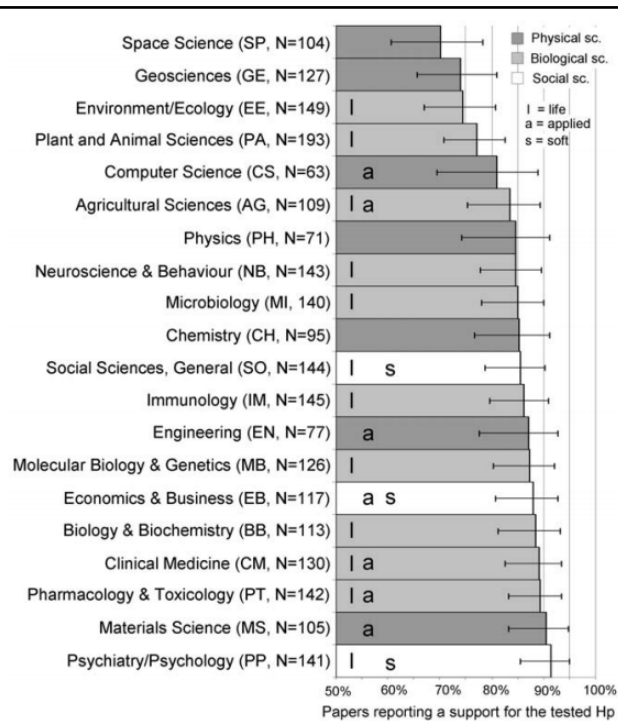
go.nature.com/nqyohl
© Nature

59



60

Fanelli, Daniele.
"Positive" results increase down the hierarchy of the sciences. PLoS One 5.4 (2010)



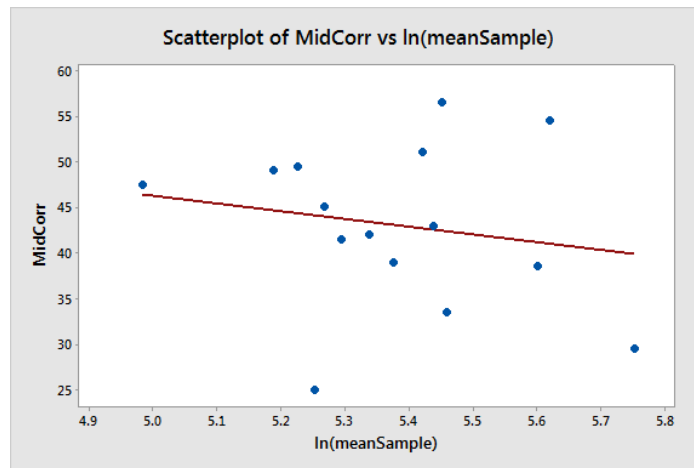
61

When are studies more likely to give incorrect results (from Ioannidis)

- Low sample size (low statistical power)
- Small (true) effect size (low statistical power, unless very large sample size)
- High the number of relationships tested, and the selective reporting (publication bias)
- High flexibility in design and interpretations, e.g., flexibility related to measures, statistical tests, study design, model tuning, definition of outliers, interpretation of data (researcher bias)
- Substantial degree of vested interests or wish for a particular outcome (researcher bias)
- Hot scientific topic (researcher bias).

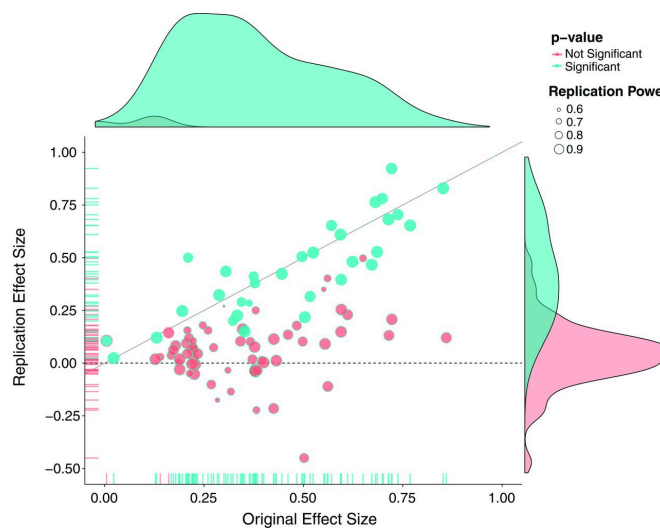
62

Schepers, Jeroen, and Martin Wetzels. "A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects." *Information & Management* 44.1 (2007): 90-103.



63

Fig. 3 Original study effect size versus replication effect size (correlation coefficients).



Open Science Collaboration *Science* 2015;349:aac4716



Published by AAAS

64

Increase the statistical power of the studies

- ✓ I see no good reason to conduct studies with power of about 40% or less for likely effect sizes. Should be at least 80%?
- **Practical consequences:**
 - ✓ Conduct a power analysis to calculate what is a sufficient number of observations.
 - ✓ If not possible to get enough observations for decent level of statistical power, then cancel the study to avoid wasting resources and to avoid getting tempted to use of questionable practises – which works much better for low power studies.
 - ✓ Do not argue that finding significant results with low power studies increases the strength of the result.

65

Introduce fewer hypotheses and improve the reporting of the results from the tests

- **Practical consequences:**
 - ✓ “Less is more”. Many tests in one study limit the value of each single test!
 - ✓ Avoid statistical tests on exploratory (post hoc) hypotheses.
 - ✓ Report on all tests, especially when they are on variants on the same dependent variable (same construct).
 - ✓ Decide as much as possible on inclusion/exclusion (outlier) criteria, statistical instruments in advance.

66

- **Improve review processes**
 - ✓ Journals and conferences should accept good studies with non-significant results.
- **More replications and meta-analyses**
 - ✓ Preferable independent replications
- **Use confidence intervals of effect sizes, rather than p-values and test of null hypotheses**
 - ✓ p-values are much too complex and much misused

67

Other possible actions:

- Protocols where hypotheses are reported before the study is conducted
- Blinding data when analysing (you should not know which one is the hypothesized direction when analysing)
- Places where non-significant results are reported
 - ✓ Journal of articles in support of the null-hypothesis exists!
- Use of Bayesian statistics
- p-value adjustments when many tests
- Better training in empirical studies and statistical methods
- Do we think any of these will work? How to make them work?

68

An example of the challenge of interpreting p-value in studies with low statistical power (which is the common situation for empirical software engineering studies)

Distribution of mean values of two treatments
Treatment B (red) is 0.25 better than Treatment A (black)

Significance level is 5% (i.e., 5% of the Treatment A distribution is right of 0.82)

Statistical power is 10% (i.e. 10% of the Treatment B distribution is right of 0.82 = the value giving $p=0.05$)

This shows that even when finding $p=0.05$ the alternative hypothesis is not much more likely than the null hypothesis!

Bayes Factor (BF) indicates knowledge increase when observing a statistically significant finding.
 $BF = \text{Likelihood of observing } p < 0.05 \text{ if true effect} / \text{likelihood of observing } p < 0.05 \text{ if no true effect} = \text{power} / \text{significance level} = 10\%/5\% = 2.0 = \text{"barely worth mentioning"}$.

69

Low power of empirical studies of SE/IS (as in many other domains) has been repeatedly documented: 1989

Table 4. Frequency and Cumulative Percentage Distribution of the Statistical Power of 57 MIS Studies*

Statistical Power Level	Small Effect		Medium Effect		Large Effect	
	Frequency	Cumulative Percentage	Frequency	Cumulative Percentage	Frequency	Cumulative Percentage
.91 - .99	—	—	40	100%	90	100%
.81 - .90	2	100%	11	73%	8	40%
.71 - .80	—	—	8	66%	11	34%
.61 - .70	2	99%	18	60%	15	27%
.51 - .60	6	97%	12	48%	11	17%
.41 - .50	5	93%	6	40%	3	9%
.31 - .40	2	90%	21	36%	7	7%
.21 - .30	30	89%	20	22%	1	3%
.11 - .20	42	68%	11	9%	2	2%
.00 - .10	60	40%	2	1%	1	1%
TOTAL	149	—	149	—	149	—
Average Power	0.19		0.60		0.83	

* Assuming small, medium, and large effect sizes, a non-directional test, and a 0.05 significance criterion.

Baroudi, Jack J., and Wanda J. Orlikowski. "The problem of statistical power in MIS research." MIS Quarterly (1989): 87-106.

70

The relation between statistical power, effect size and significance levels

	EFFECT=TRUE	EFFECT=FALSE
Significant result for test of hypothesis ($p\text{-value} > \alpha$)	TRUE POSITIVE Claiming an effect that is there. (Correct result)	FALSE POSITIVE Claiming an effect that is not there. (Incorrect result)
Non-significant result for test of hypothesis ($p\text{-value} \leq \alpha$)	FALSE NEGATIVE Not finding an effect that is there. (Incorrect result)	TRUE NEGATIVES Not claiming an effect that is not there. (Correct result)

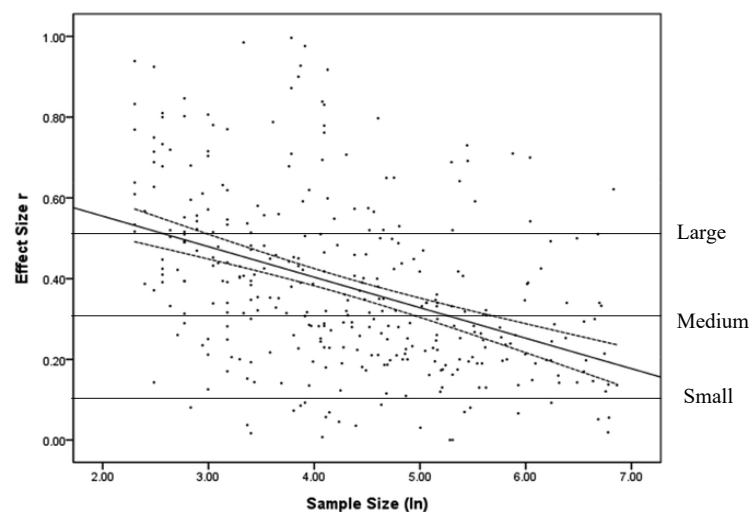
Effect size: The strength (size) of the effect. Examples of effect size measures: Correlation, Odds ratio, Cohen's d, Percentage difference.

Statistical power: Probability of $p \leq \alpha$, if there is a true effect (for a given effect size).

p-value: The probability of observing the data (or more extreme data), given that there is no effect, i.e., $p(D | H_0)$.

71

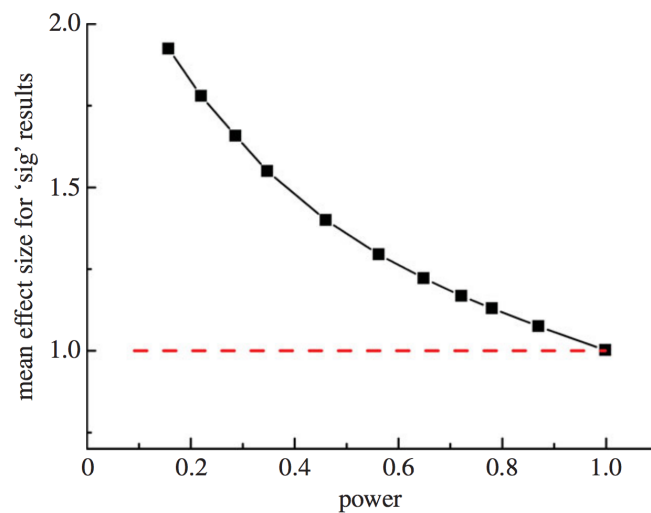
Figure 5. Corrected effect size r plotted against logarithmically transformed sample size.



Kühberger A, Fritz A, Scherndl T (2014) Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. PLoS ONE 9(9): e105825. doi:10.1371/journal.pone.0105825
<http://dx.doi.org/10.1371/journal.pone.0105825>

72

Relation between effect size and statistical power when publishing only statistically significant results
(and true effect is 1.0)



73

A brief side-track on p-values

a p-value around 0.05 is often a weak result – especially when the statistical power is low - leading to low result reliability

74

p-values are complex, unreliable, misunderstood values that do not answer what we should be asking about ... (and part of the result reliability problem!)

A p-value is **not** the probability of the null hypothesis (or alternative hypothesis) being true! A p-value of 0.05 may frequently correspond to a much higher probability that the null hypothesis is true.

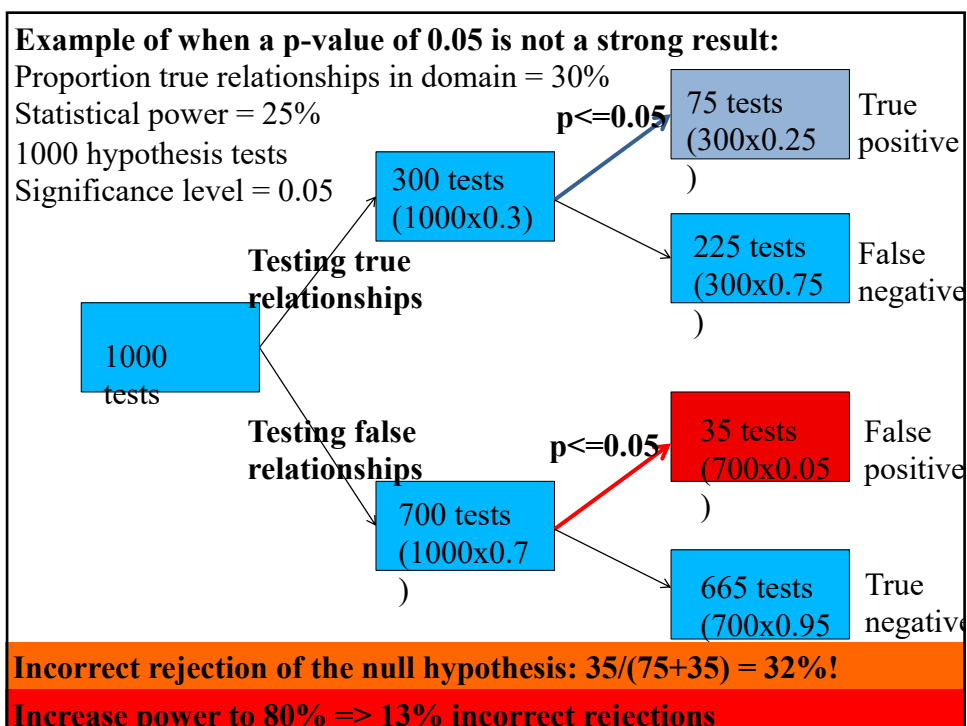
A p-value does **not** tell how likely it is to replicate the study and find $p < 0.05$, e.g., that repeating the study 100 times would result in 95 being statistically significant. (Same sample size, $p = 0.05$ and true effect size, means only 50% likely to replicate. Replications of findings with $p = 0.05$ should typically more than double the sample size to have a reasonable probability of finding $p < 0.05$)

Even with $p = 0.05$, the null hypothesis may be more likely than the alternative hypothesis (e.g., when the statistical power is very low)

The p-value examines a “yes/no” situation, while we in most cases would like to know about the effect size and its uncertainty.

We should start using confidence intervals of effect sizes, rather than p-values.

75



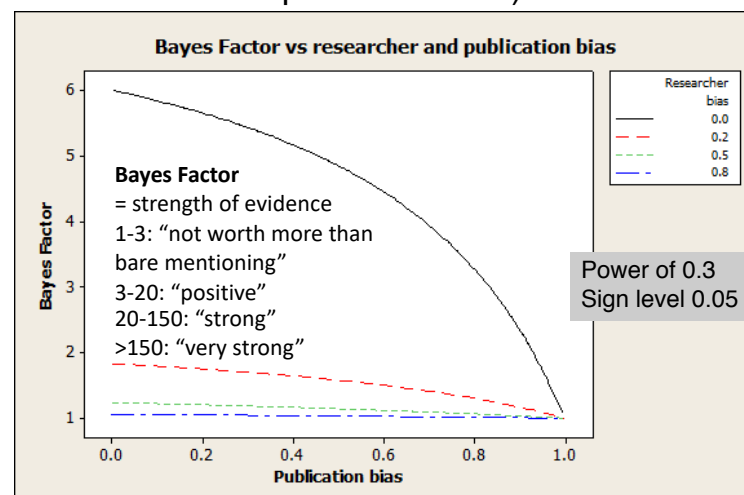
76

A P-VALUE < 0.05 IS
 CONSEQUENTLY FAR FROM A
 GUARANTEE FOR A RELIABLE
 RESULT WHEN THE STATISTICAL
 POWER IS LOW

(EVEN WITHOUT ANY
 RESEARCH AND PUBLICATION
 BIAS!)

77

The Bayesian way of looking at this ...
 (shows the low value in studies with low power + research
 bias + publication bias)



78