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

Agenda of the workshop

- Results on the state-of-reliability of empirical results in software engineering. (30 minutes)
  - **Magne Jørgensen**
- Responses and reflections from the panel. (30 minutes)
- Panel members:
  - **Natalia Juristo/Sira Vegas**
  - **Maurizio Morisio**
  - **Günter Ruhe (new EiC for IST)**
- Discuss the following questions with you (30 minutes):
  - How bad is the situation? How much can we trust the results?
  - What should we do? What are *realistic*, *practical* means to improve the reliability of empirical software engineering results?
- PS: The question of **industry impact** is also an important issue, but maybe for another workshop.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

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 preselection 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 on

is characte
vary 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 also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is 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 c 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

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

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

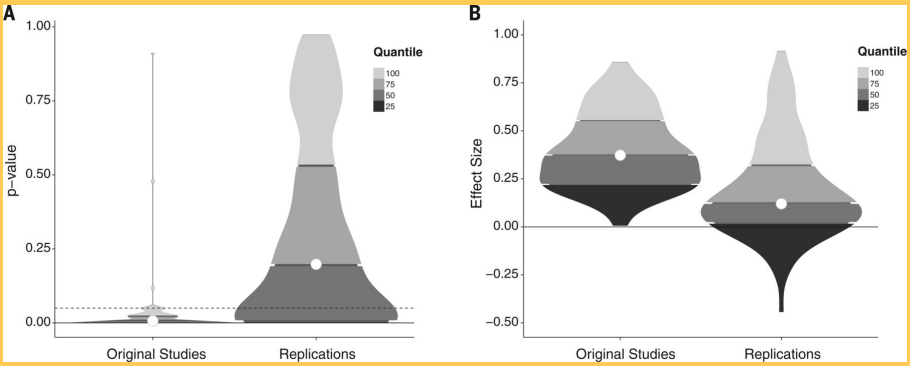
Nature, October 2015, Regina Nuzzo

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.

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?

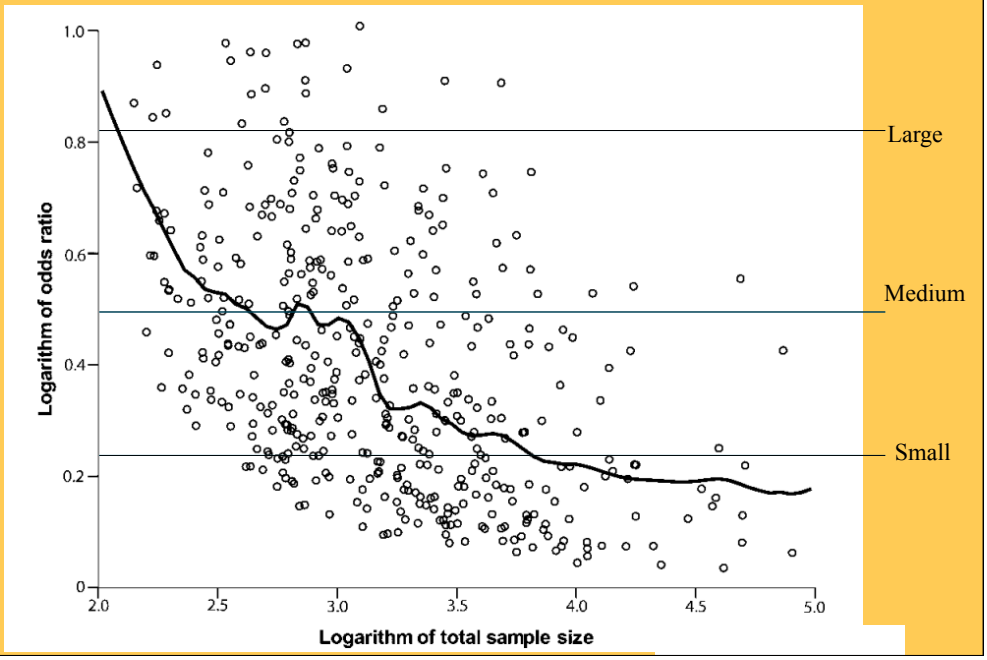
OUR STUDY INDICATES THAT WE WILL FIND SIMILARLY DISAPPOINTING RESULTS IN EMPIRICAL SOFTWARE ENGINEERING

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.

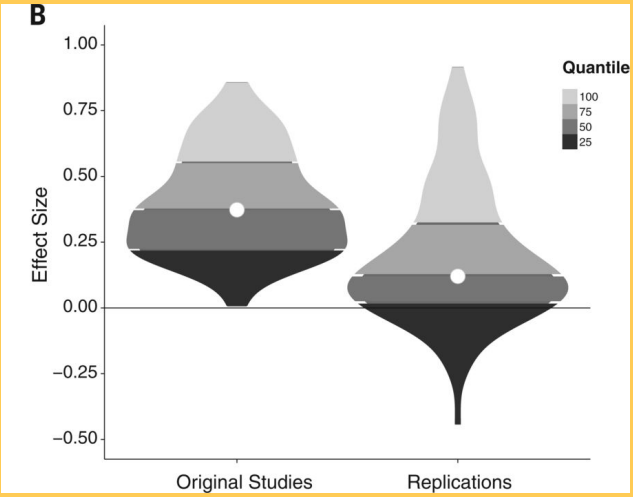
EXAGGERATED EFFECT SIZES OF SMALL STUDIES

"Why most discovered true associations are inflated", Ioannidis, Epidemiology, Vol 19, No 5, Sept 2008



PSYCHOLOGY: Decrease from medium (correlation = 0.35) to low (correlation = 0.1) effect size in replicated studies with high statistical power.

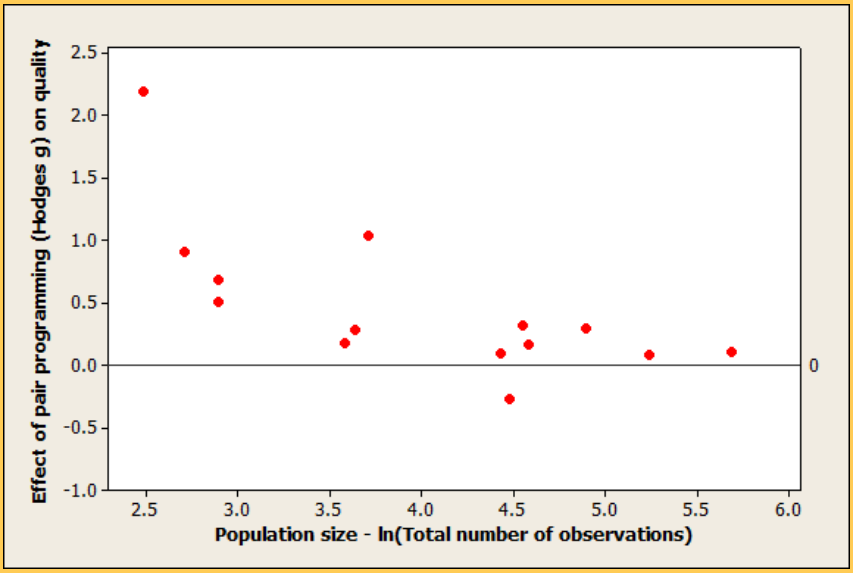
Reference: Open Science Collaboration. Estimating the reproducibility of psychological science. Science 349.6251 (2015): aac4716.



Difficult to predict which of the studies where they would be able to replicate the original result!

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.



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).**

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

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

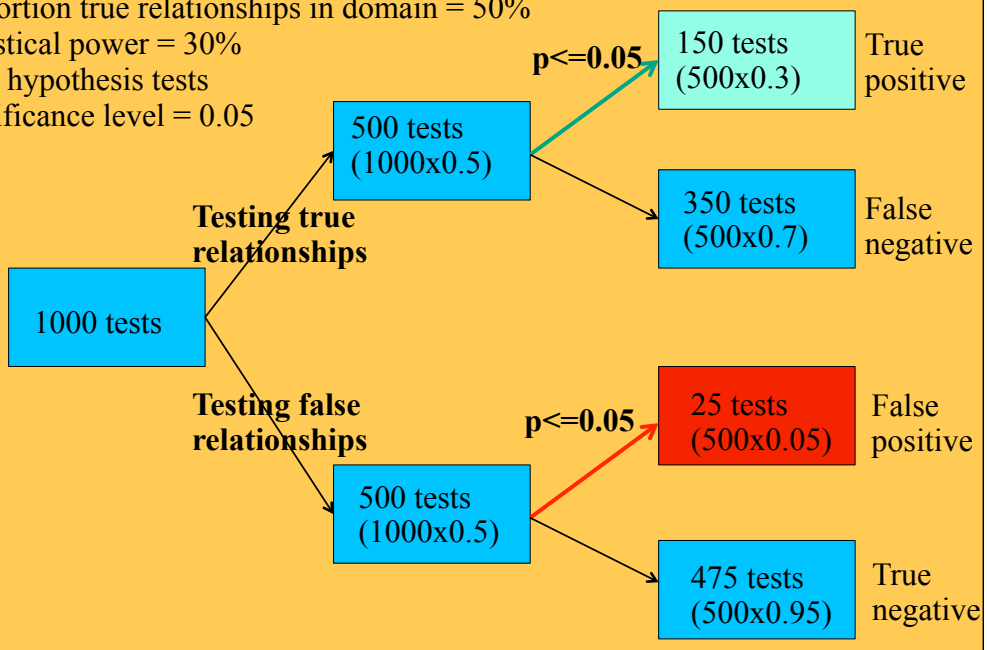
Power level	Small effect size		Medium effect size		Large effect size	
	Freq.	Cum. %	Freq.	Cum. %	Freq.	Cum. %
.91–.99	–	–	18	100	69	100
.81–.90	1	100	11	96	75	85
.71–.80	–	100	14	94	49	69
.61–.70	2	100	13	91	70	58
.51–.60	9	99	44	88	58	43
.41–.50	2	97	50	78	21	30
.31–.40	–	97	76	67	43	25
.21–.30	13	97	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.

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.

Example: Proportion of statistically signifcant findings
Proportion true relationships in domain = 50%
Statistical power = 30%
1000 hypothesis tests
Significance level = 0.05



Expected statistically significant relationships: (25+150)/1000 = 17.5%

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

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%

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?

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%

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.

Practices leading to research and publication bias

TABLE 1: Results from a survey on statistical practices

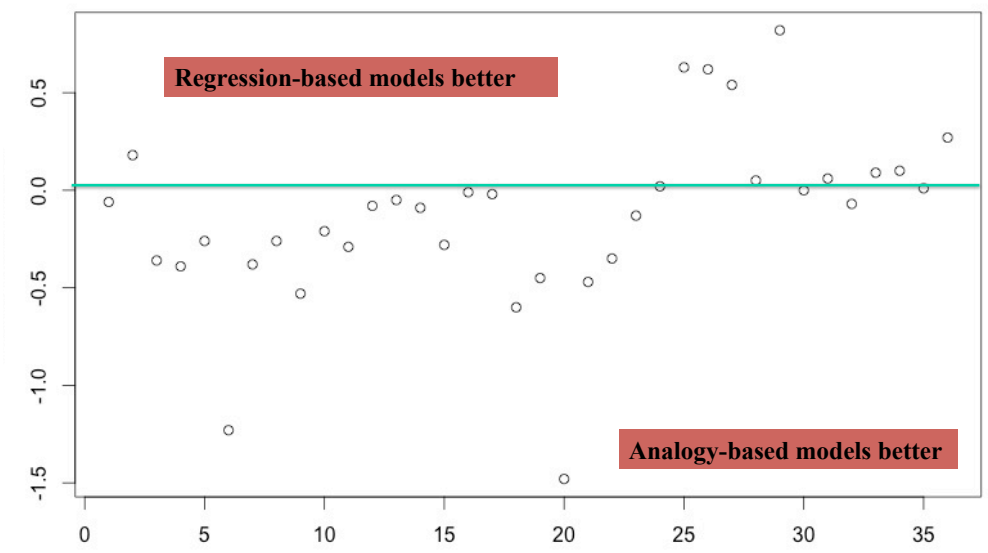
Research Practice	Have experienced/done this in my own research				
	Never	Seldom	Occasionally	Often	Don't know
P1: Paper rejected due to non-significance[1]	14	6	8	4	4
P2: Paper not submitted due to non-significance[2]	16	6	8	4	1
P3: Not reported non-significant results[3]	17	8	4	4	2
P4: Not reported 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."*

HOW MUCH RESEARCHER BIAS IS THERE?

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

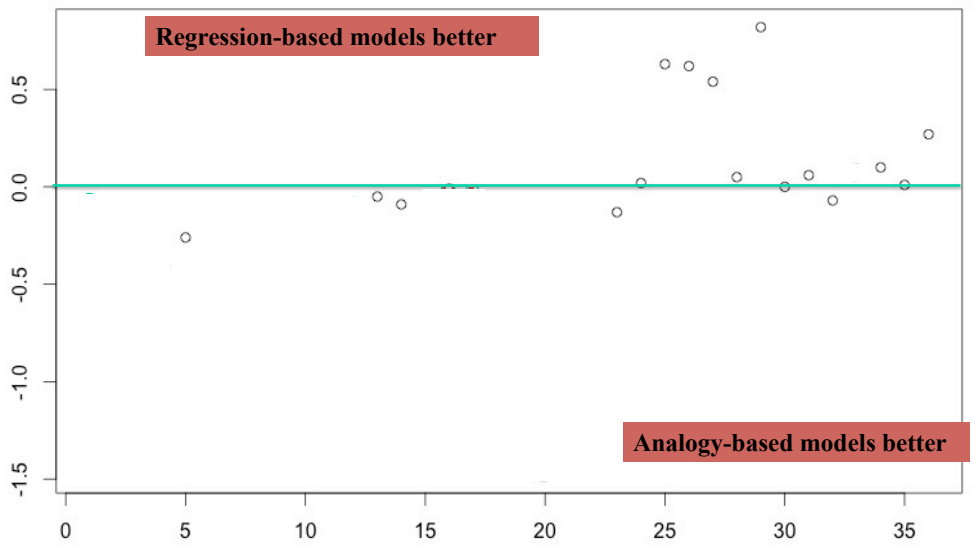
Effect size = MMRE_analogy – MMRE_regression



All studies: Analogy-based estimation models are typically more accurate

Effect size = MMRE_analogy – MMRE_regression

Removed studies evaluating own model (vested interests, likely research bias)



Neutral studies: Regression-based estimation models are typically more accurate

AN ILLUSTRATION OF THE EFFECT OF A LITTLE RESEARCH AND PUBLICATION BIAS:

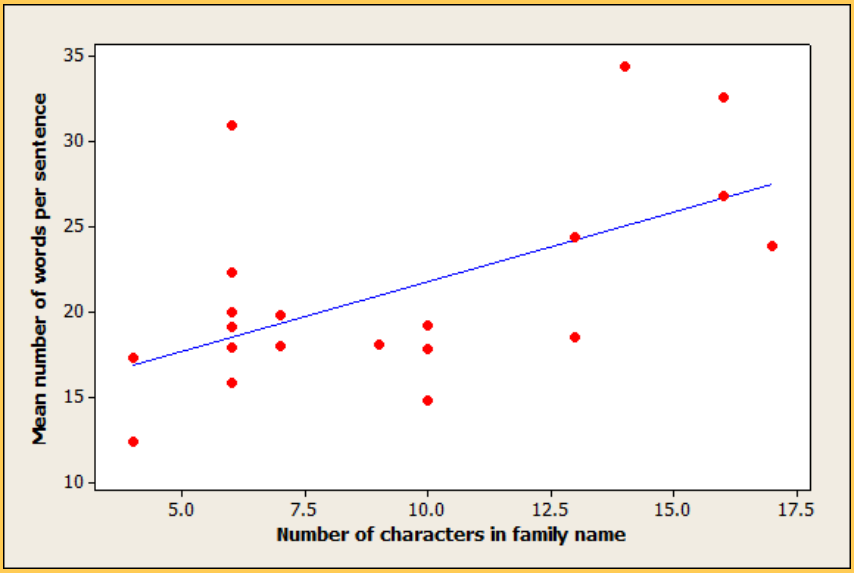
You should try something like the following experiment yourself – either with random data, or with "silly hypotheses" – to experience how easy it is to find $p<0.05$ with low statistical power and some questionable, but common practices.

My hypothesis: People with longer names write more complex texts

Heureka! p<0.05 & medium effect size

- **Variables**:
  - LengthOfName: Length of surname of the first author
  - Complexity1: Number of words per paragraph
  - Complexity2: Flesch-Kincaid reading level
- **Correlations**:
  - $r_{LengthOfName,Complexity1} = 0.581$ (p=0.007)
  - $r_{LengthOfName,Complexity2} = 0.577$ (p=0.008)
- **Data collection**:
  - The first 20 publications identified by “google scholar” using the search string “software engineering”.

A regression line supports the results



How did I do it?

(How to easily get p<0.05 in any low power study)

- **Publication bias**: Only the two significant, out of several tested, 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 effect size of r=0.3, p<0.05).
  - A significant effect with low power is NOT better than one with high power – although several researchers make this claim

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 reproducable and which not.
- We need less ”shotgun” type of hypthesis testing and more hypotheses based on theory and prior explorations (”less is more” when it comes to hypothesis testing)

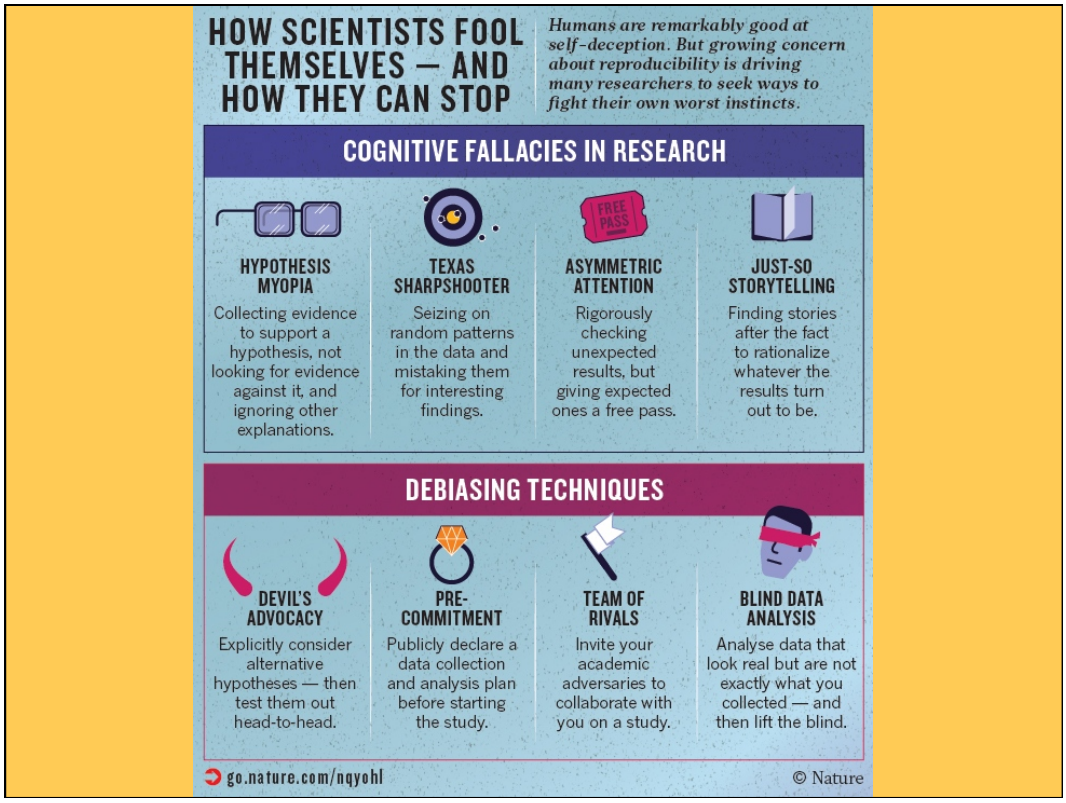
Questions to discuss

- Is the situation as bad it looks like?
  - How big is the problem in practice?
  - Are there contexts – types of studies - we can trust much more than others?
- What are *realistic*, *practical* means to improve the reliability of empirical software engineering?
  - What is the role of editors and reviewers to improve the reliability situation?
- What has stopped us from improving so far? We have known about most of the problems for quite some time.
- Are there good reasons to be optimistic about the future of empirical software engineering?

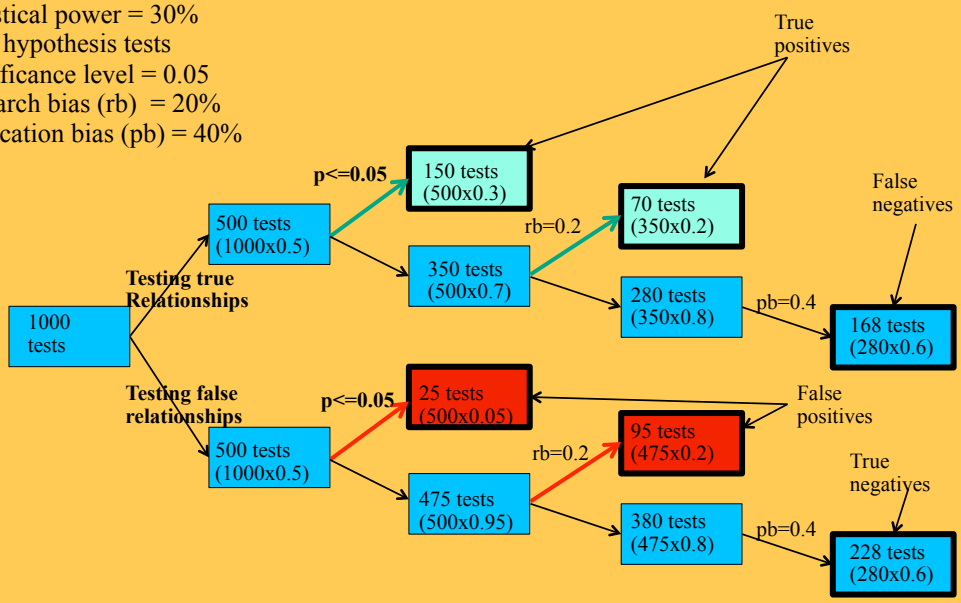
Agenda of the workshop

- Results on the state-of-reliability of empirical results in software engineering. (30 minutes)
  - **Magne Jørgensen**
- Responses and reflections from the panel. (30 minutes)
- Panel members:
  - **Natalia Juristo/Sira Vegas**
  - **Maurizio Morisio**
  - **Günter Ruhe (new EiC for IST)**
- Discuss the following questions with you (30 minutes):
  - How bad is the situation? How much can we trust the results?
  - What should we do? What are *realistic*, *practical* means to improve the reliability of empirical software engineering results?
- PS: The question of **industry impact** is also an important issue, but maybe for another workshop.

EXTRA



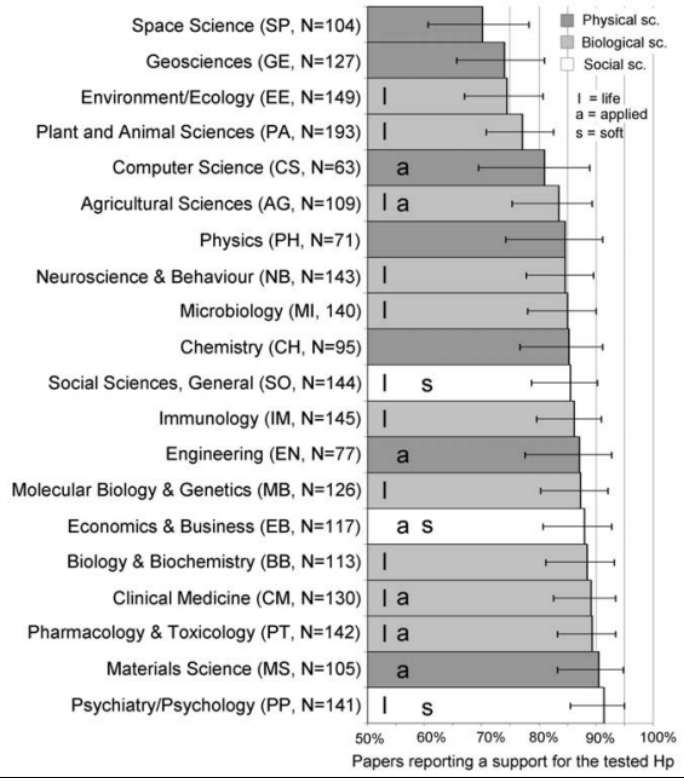
Example adding research and publication bias:
Proportion true relationships in domain = 50%
Statistical power = 30%
1000 hypothesis tests
Significance level = 0.05
Research bias (rb) = 20%
Publication bias (pb) = 40%



Proportion statistically significant results: (150+70+25+95)/736=46%

Incorrect statistically significant results: (25+95)/ (150+70+25+95) = 32%

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



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).

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.

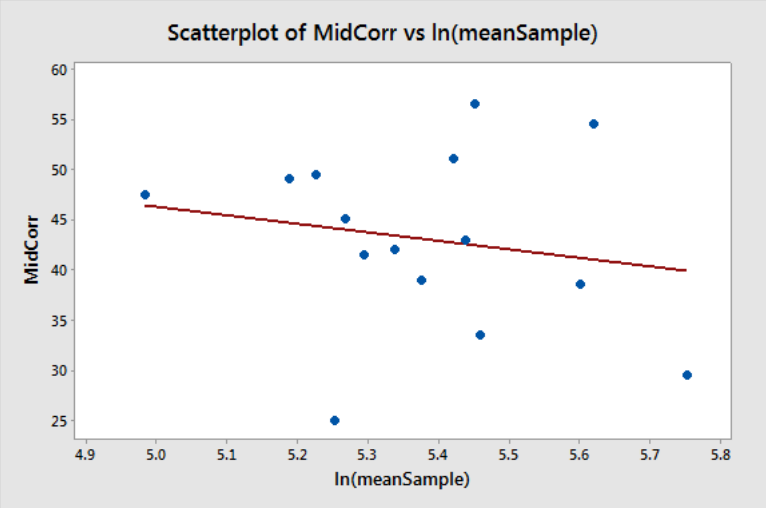
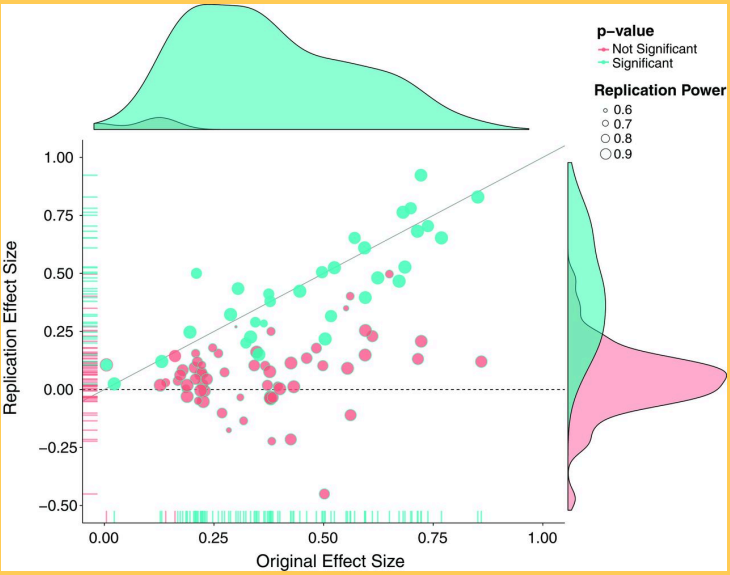


Fig. 3 Original study effect size versus replication effect size (correlation coefficients).

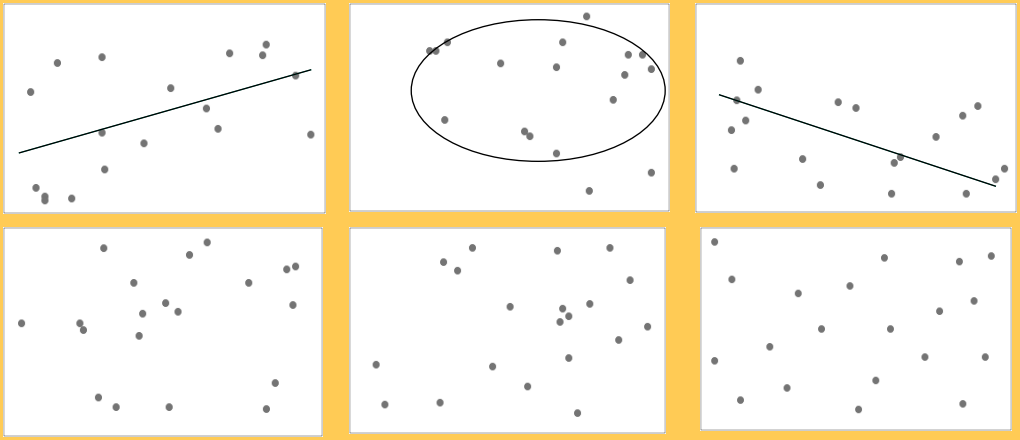


Open Science Collaboration Science 2015;349:aac4716

Science
AAAS

Published by AAAS

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 generated from a random data generator.)

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.

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.

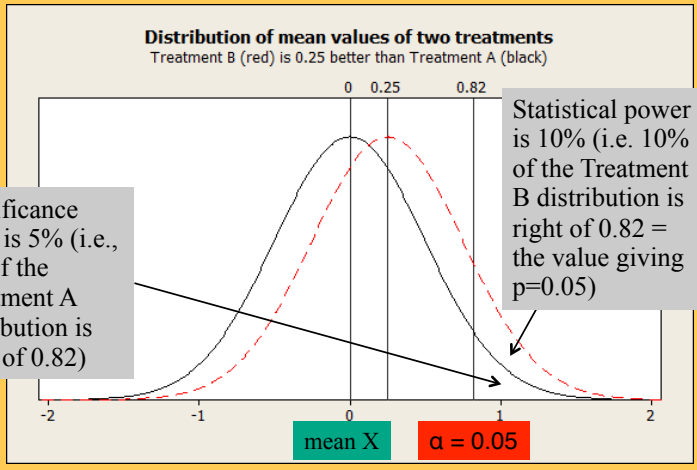
- **Improve review processes**
  - Journals and conferences should accept good studies with non-significant results.
- **More replications and meta-analyses**
  - Preferable independent replications
- **Use confidence intervals of effect sizes, rather than p-values and test of null hypotheses**
  - p-values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

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?

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)



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 = Likelihood of observing p<0.05 if true effect / likelihood of observing p<0.05 if no true effect = power / significance level = 10%/5% = 2.0 = "barely worth mentioning".

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*

Stasticial 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.

The relation between statistical power, effect size and significance levels

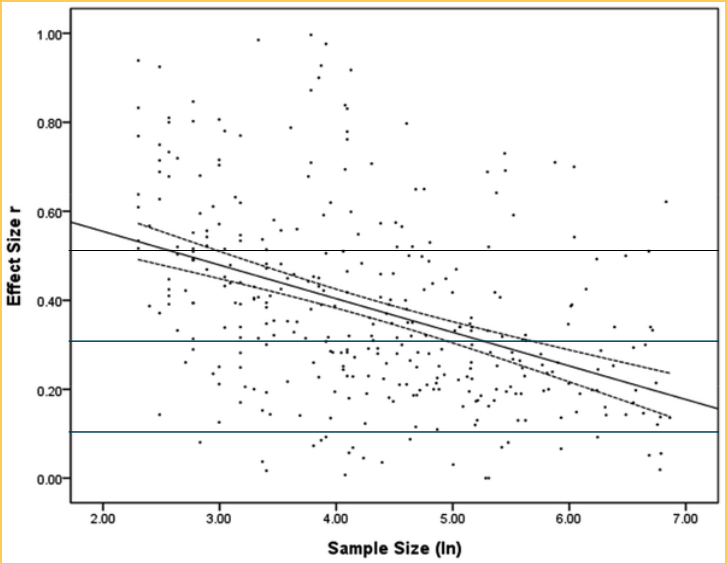
	EFFECT=TRUE	EFFECT=FALSE
Significant result for test of hypothesis (p-value > α)	**TRUE POSITIVE** Claiming an effect that is there. (Correct result)	**FALSE POSITIVE** Claiming an effect that is not there. (Incorrect result)
Non-significant result for test of hypothesis (p-value <= α)	**FALSE NEGATIVE** 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<=α, 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., $p(D \mid H_0)$.

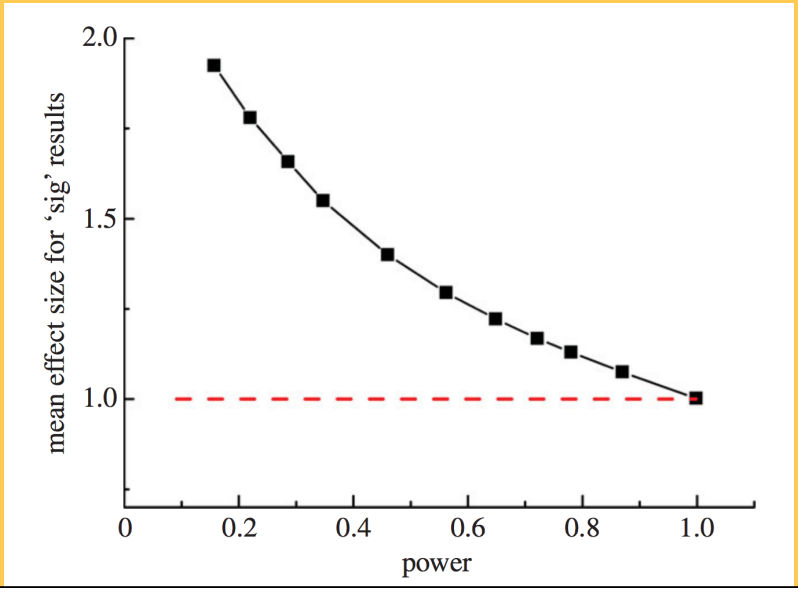
Figure 5. Corrected effect size r plotted against logarithmically transformed sample size.



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://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0105825

PLOS ONE

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



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

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.*

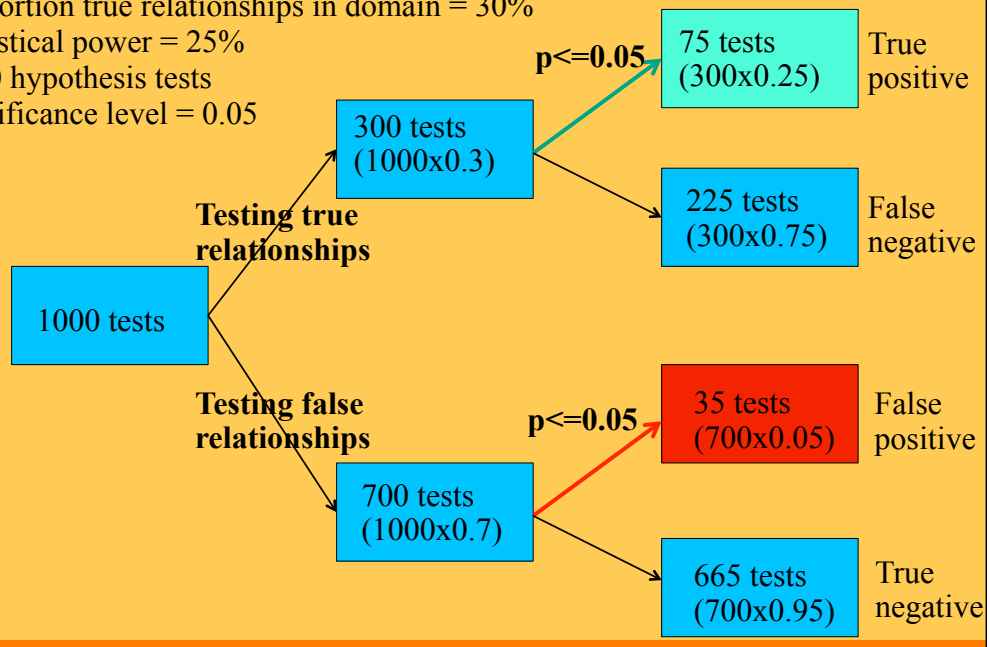
*A p-value does **not** tell how likely it is to replicate the study and find p<0.05, e.g., that repeating the study 100 time would result in 95 being statistically significant. (Same sample size, p=0.05 and true effect size, means only 50% likely to replicate. Replications of findings with p=0.05 should 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.

Example of when a p-value of 0.05 is not a strong result:
Proportion true relationships in domain = 30%
Statistical power = 25%
1000 hypothesis tests
Significance level = 0.05



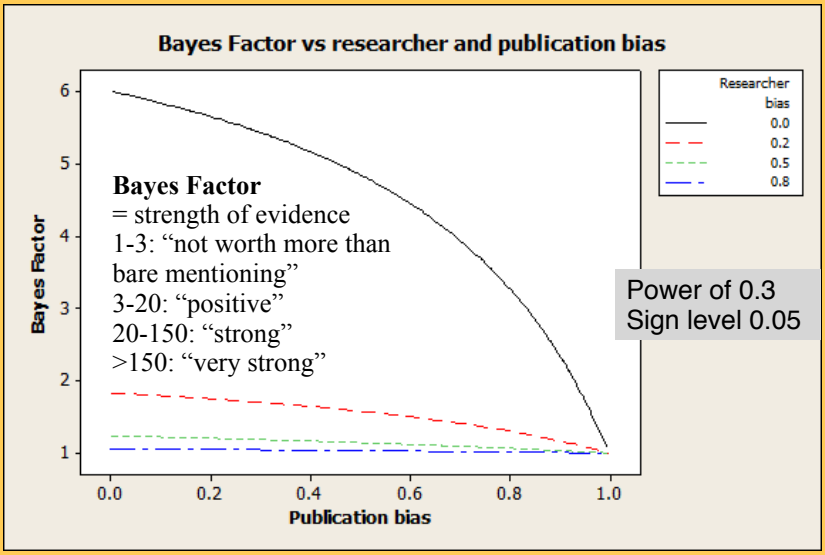
Incorrect rejection of the null hypothesis: 35/(75+35) = 32%!

Increase power to 80% => 13% incorrect rejections

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!)

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



Power of 0.3
Sign level 0.05