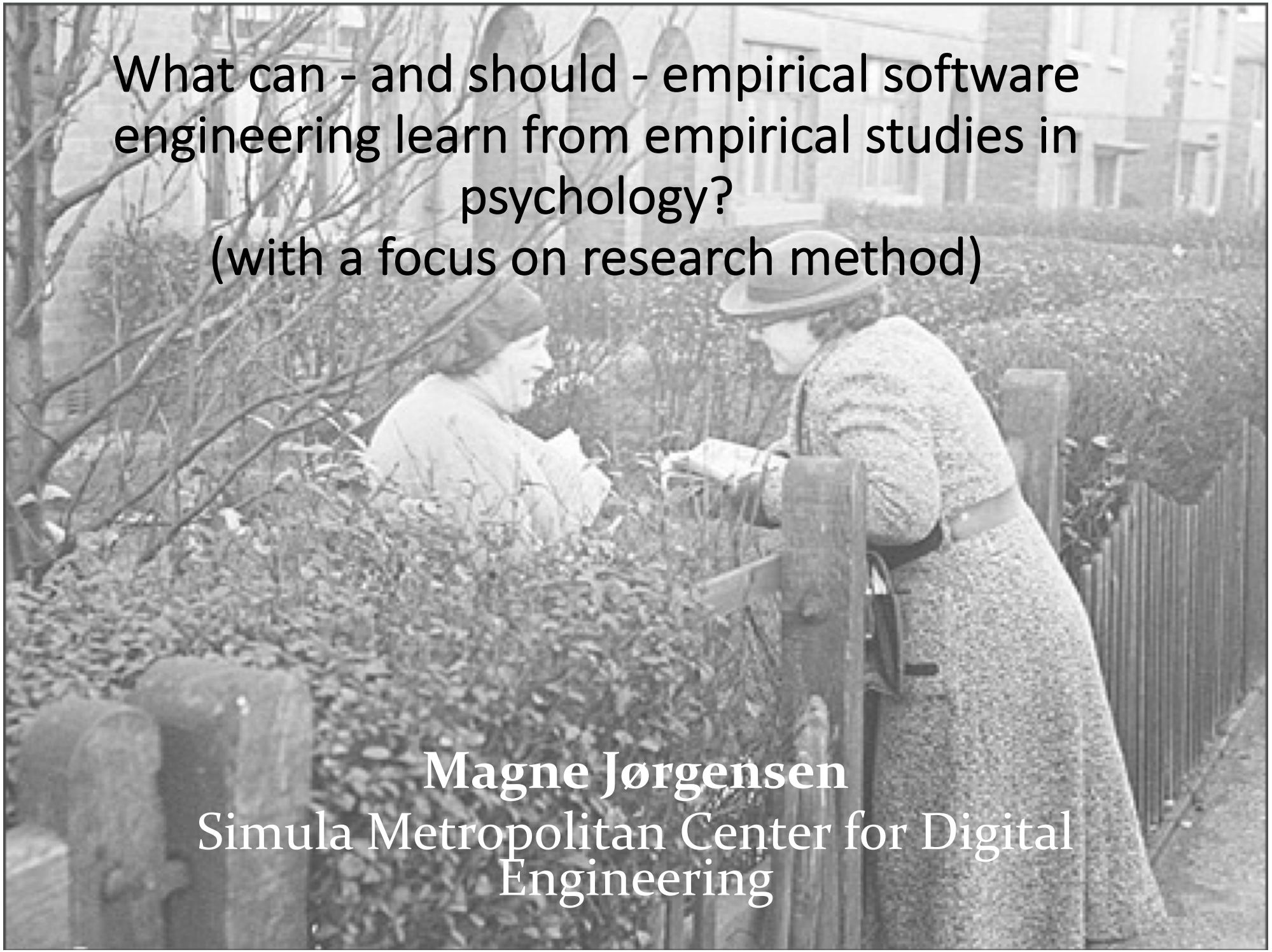


What can - and should - empirical software engineering learn from empirical studies in psychology?
(with a focus on research method)

Magne Jørgensen
Simula Metropolitan Center for Digital Engineering

Why I think we can and should learn from empirical psychology

We share the interest in human behaviour and decision making

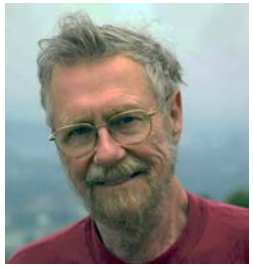
Empirical psychology is a much older and larger discipline

The first psychology lab with empirical studies was established in the 1870s.

A short side-track (before we go back to what we can learn)

How large role does psychology research play in software engineering research?

Pre-empirical software engineering: *”Go to considered harmful”*



- *Edgar Dijkstra: “Go to statement considered harmful”. Communications of ACM. 1968.*
- *Donald Knuth: Structured programming with go to statements. ACM Computing Surveys, 1974.*
- *Two brilliant men, discussing the topic by giving tailored examples of bad and good use of GOTOs. Many organizations banned the use of GOTOs.*
- *The topic has been discussed a lot since then: Stack Overflow introduced a topic in 2008 titled “GOTO still considered harmful?”. This has been viewed about 60.000 times.*
- *First empirical study on GOTOs not before 2015! Nagappan, Meiyappan, et al. "An empirical study of goto in C code from GitHub repositories.” Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. ACM, 2015.*

GOTOs: The empirical evidence

- Nagappan, Robbes, Kamei, Tanter, McIntosh, Mockus and Hassan found through an analysis of more than 10.000 projects and 2 million C-files:
  - Much use of GOTOs. Especially for systems and networks projects. More than one fourth of the analysed projects used GOTOs.
  - GOTOs mainly used for error handling and cleaning up resources at the end of a procedure.
  - Programmers did hardly ever create ”spaghetti code” with GOTOs.
- Dijkstra’s worries were not empirically justified then - that is of course not to say that he could have been right - and not now.
- Not much harm in continued use GOTOs, as long as it goes along with as much care and smartness as before.

Early pioneers with research methods inspiration from natural sciences

- L. A. Belady and M. M. Lehman, *Programming Systems Dynamics, or the Meta-Dynamics of Systems in Maintenance and Growth*, IBM Research Report RC 3546 (September 1971), IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.
- L. A. Belady, A. Laszlo A., and M. M. Lehman. "*A model of large program development." IBM Systems journal* 15.3 (1976): 225-252.
  - ”Starting with the available data, we have attempted to deduce the nature of consecutive releases of os/360.”
  - “Law of continuing change: A system that is used undergoes continuing change until it is judged more cost effective to freeze and recreate it.”
  - “Law of entropy. The entropy of a system (its unstructuredness) increases with time, unless specific work is executed to maintain or reduce it.”

The software engineering laboratory at NASA

(Also much inspired by research methods in natural sciences.)

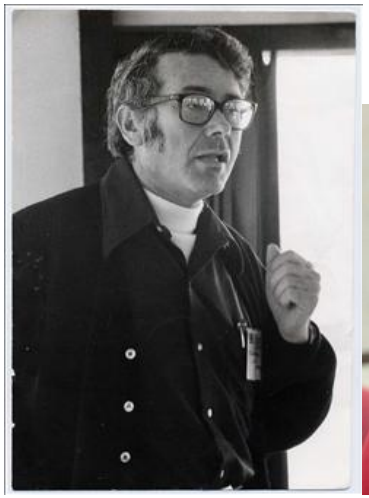
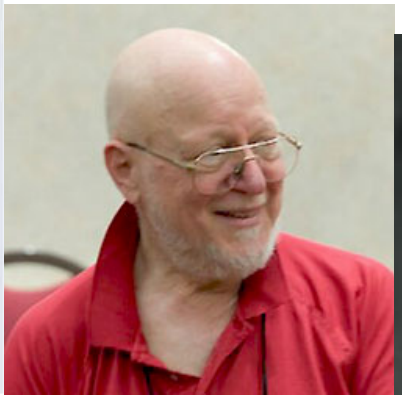
Basili, Victor R., and Marvin V. Zelkowitz. "The software engineering laboratory: Objectives." *Proceedings of the fifteenth annual SIGCPR conference*. ACM, 1977.

In a second controlled experiment, several large scale projects (5 to 10 man years each) are to be carefully monitored with some of the personnel given a training course and set methodology to use. Using the notation above, these will be a set of C projects with no A and B. While the projects are not identical, they are highly similar and should yield information about differences in techniques. In Section V, both of

Identified factors with impact on software development:

- People factors
- Problem factors
- Process factors
- Product factors
- Resource factors
- Tool factors

Even earlier pioneers with research method inspiration from psychology


- Grant, Eugene and Sackman, Harold. (1967). "*An exploratory investigation of programmer performance under on-line and off-line conditions*." LE.E.E. Transactions on Human Factors in Electronics, HFE-8.
- Weinberg, Gerald M., and Edward L. Schulman. "*Goals and performance in computer programming*." *Human factors* 16.1 (1974): 70-77.
- Shneiderman, B., Mayer, R., McKay, D., & Heller, P. (1977). *Experimental investigations of the utility of detailed flowcharts in programming*. *Communications of the ACM*, 20(6), 373-381.

Much psychology in the beginning of empirical SE and then a decline?

- Before 1990, the psychology-inspired empirical research on software engineering was relatively large (30-50% of the research?) and, I think, methodologically strong:
  - Mature use of theories from other fields
  - Good at focusing on understanding core mechanisms (mainly on problem solving)
  - Good at using industry contexts for their studies, which very often were based on controlled experiments
- What happened?
  - "Gravity" towards what we have learned? Most of us come from computer science and have little background in psychology.
  - Software engineers not as interesting as study objects as before?
  - Other reasons?

Interdisciplinarity with empirical psychology has been a success factor for much influential research

Examples:

- Herbert Simon (computer science, AI, economy, political science, organizations, cognitive psychology)
- Daniel Kahneman (economy, psychology)
- Richard Thaler (economy, psychology)

All of them received the Nobel Prize in economy for influential work. Herbert Simon also received the ACM Turing Award.

We use the same empirical methods

Similarity in research challenges – but ahead of us in awareness and dealing with them

Generalizability/transferability challenge: Very high context variability. (Ten context variables with five levels each give around 5^{10}= ca. 10 million different contexts.)

Reversibility challenge: People are adaptive, with flexible behaviour. Knowing the results/learning may affect future behaviour and thinking.

Reproducibility/replication challenge (or even crisis): Far too many results are not reproducible, and those who do have much lower effect sizes than the original study.

The reproducibility/replication crisis

The next slides are very much about the problems in quantitative studies, but there is (I think) no reason to believe that qualitative studies are better off.

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 fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the

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 artic

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

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g.,

is characte
vary a lot depending on whether th
field targets highly likely relation ships
or searches for only one or a f w
true relationships among th usands
and millions of hypothe s that may
be postulated. Let us a so consider,
for computational s mplicity,
circumscribed fi ds where either there
is only one tru relationship (among
many that c n be hypothesized) or
the powe s similar to find any of the
several existing true relationships. The
pre-study probability of a relationship
b ng 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
is the positive predictive value, PPV.
The PPV is also the complementary

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

ELSEVIER

Journal of Systems and Software

Volume 116, June 2016, Pages 133-145

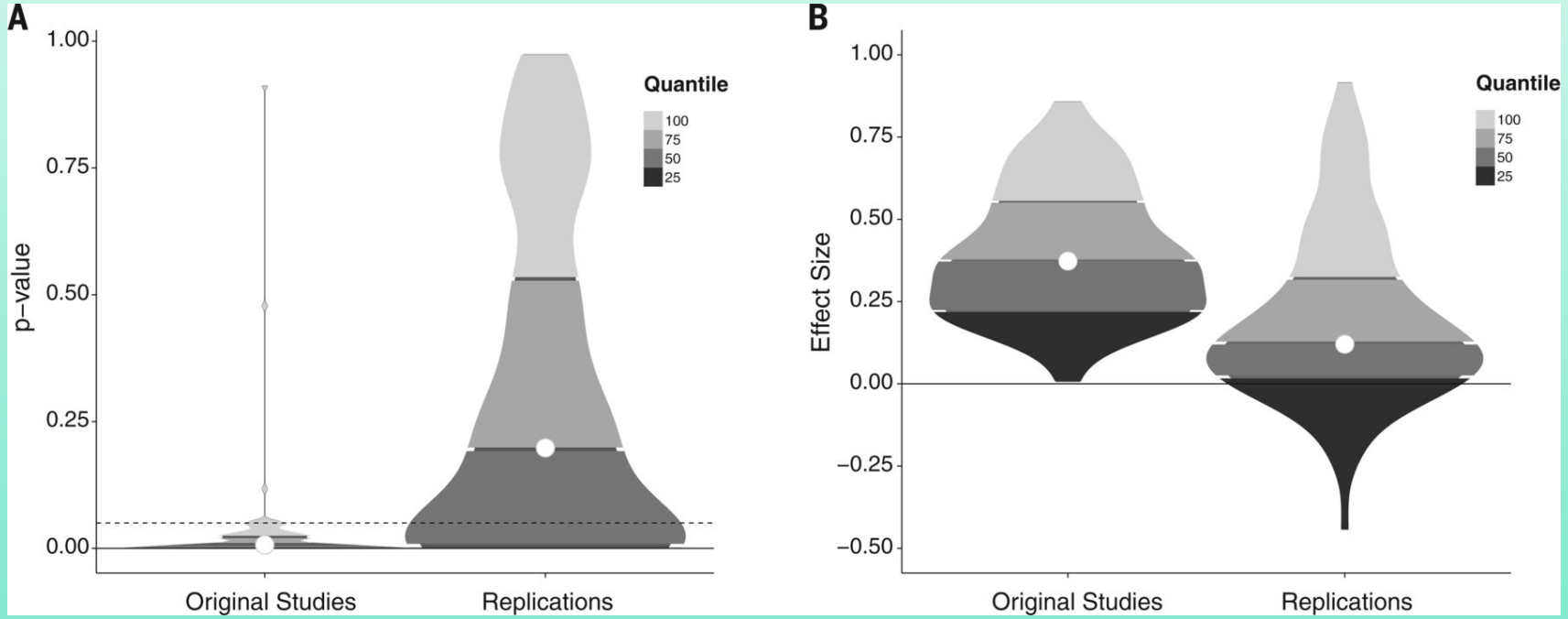
Incorrect results in software engineering experiments: How to improve research practices

Magne Jørgensen [a, b], Tore Dybå [b, c], Knut Liestøl [b], Dag I.K. Sjøberg [b]

Assume	Incorrect results	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)

The study also – perhaps more importantly – shows that there must be a large amount of researcher and publication bias in our studies

Replication of 100 experiments reported in papers published in 2008 in three top psychology journals (Replication sample size 3-4 times the original size)



Open Science Collaboration. *Estimating the reproducibility of psychological science*. Science 349.6251 (2015).

Reproduced effect size was on average about one third of the originally reported effect size.

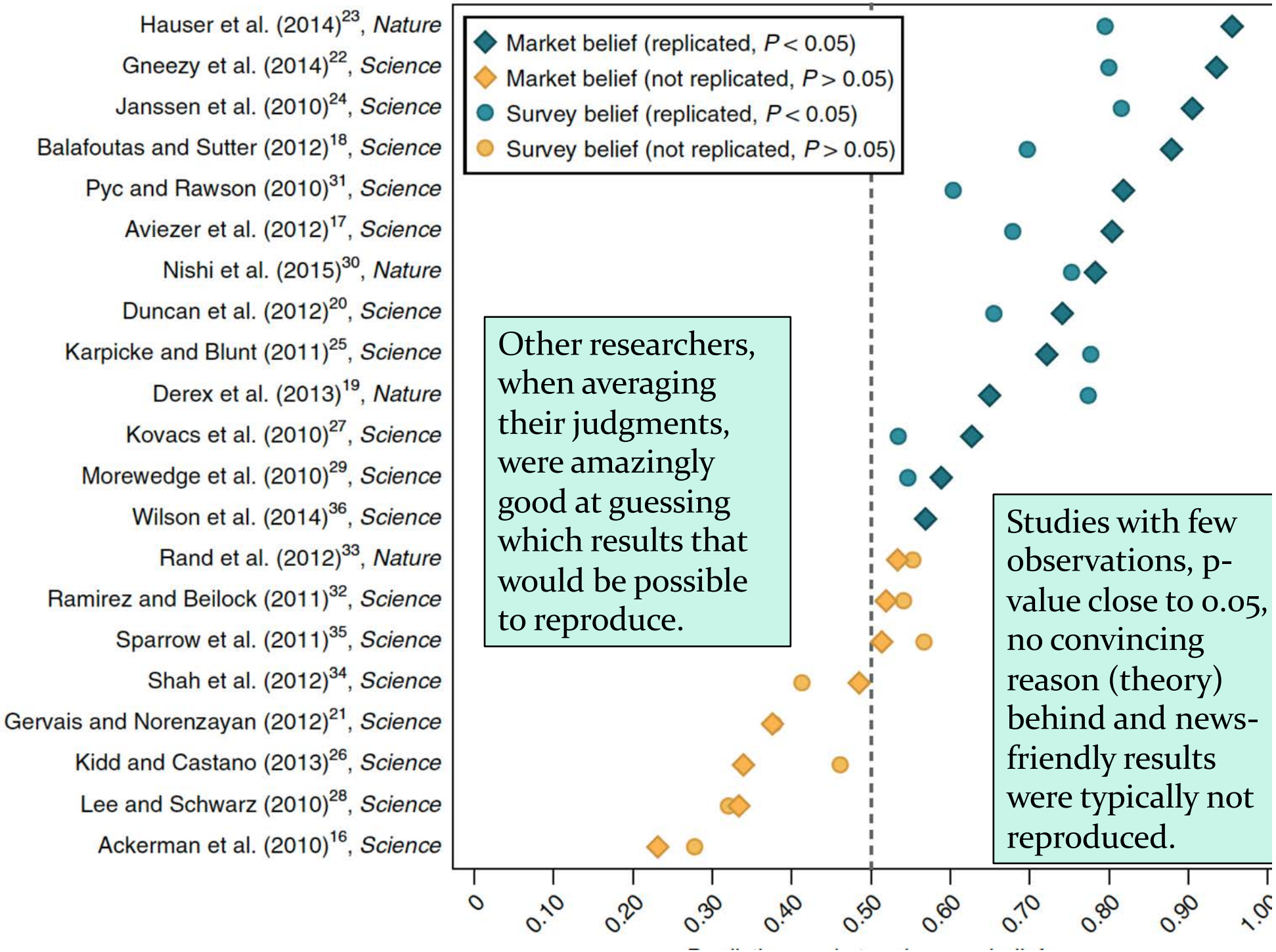
nature human behaviour

LETTERS

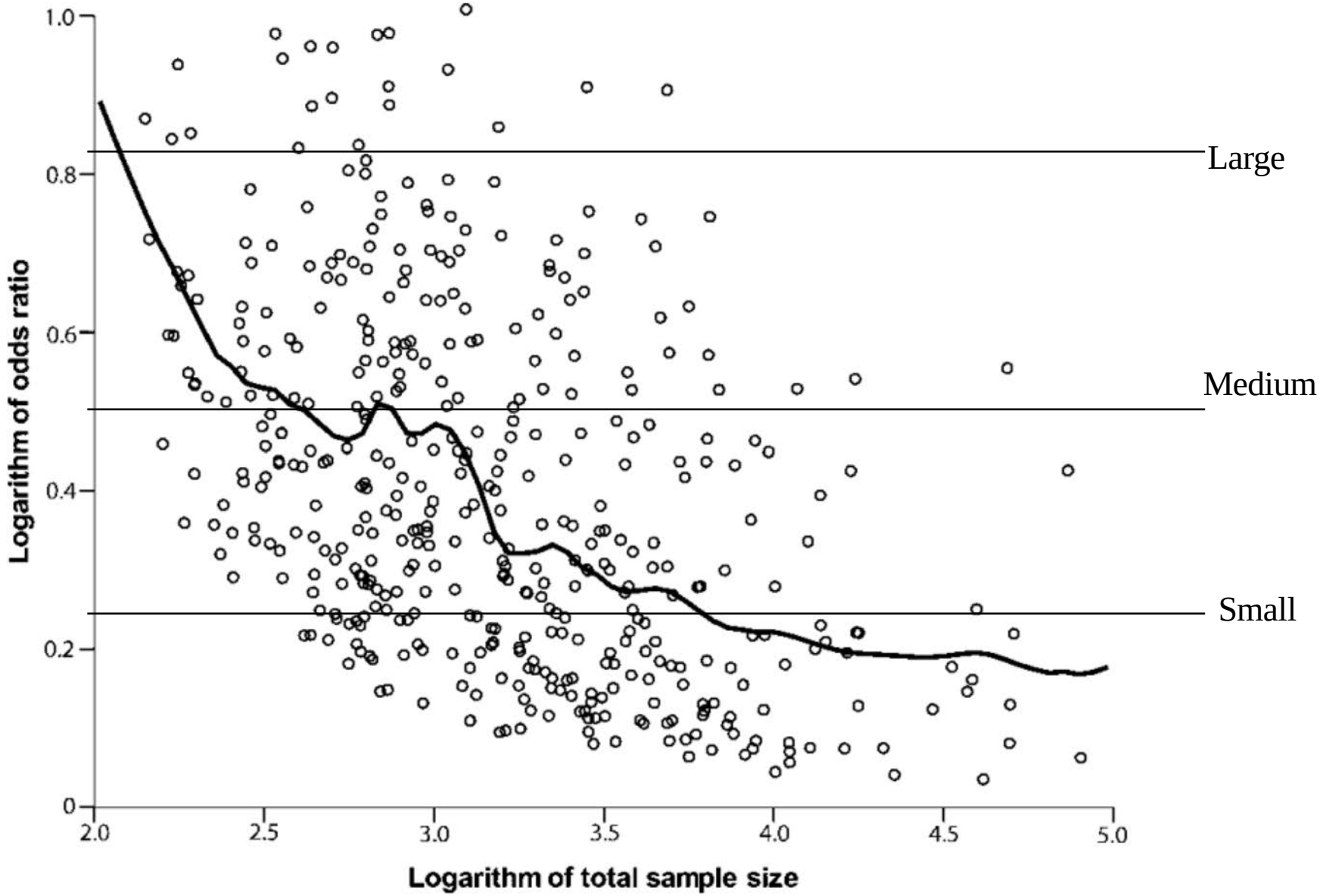
https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

- Sample sizes on average about five times higher than in the original studies.
- Statistically significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications was on average about 50% of the original effect size.

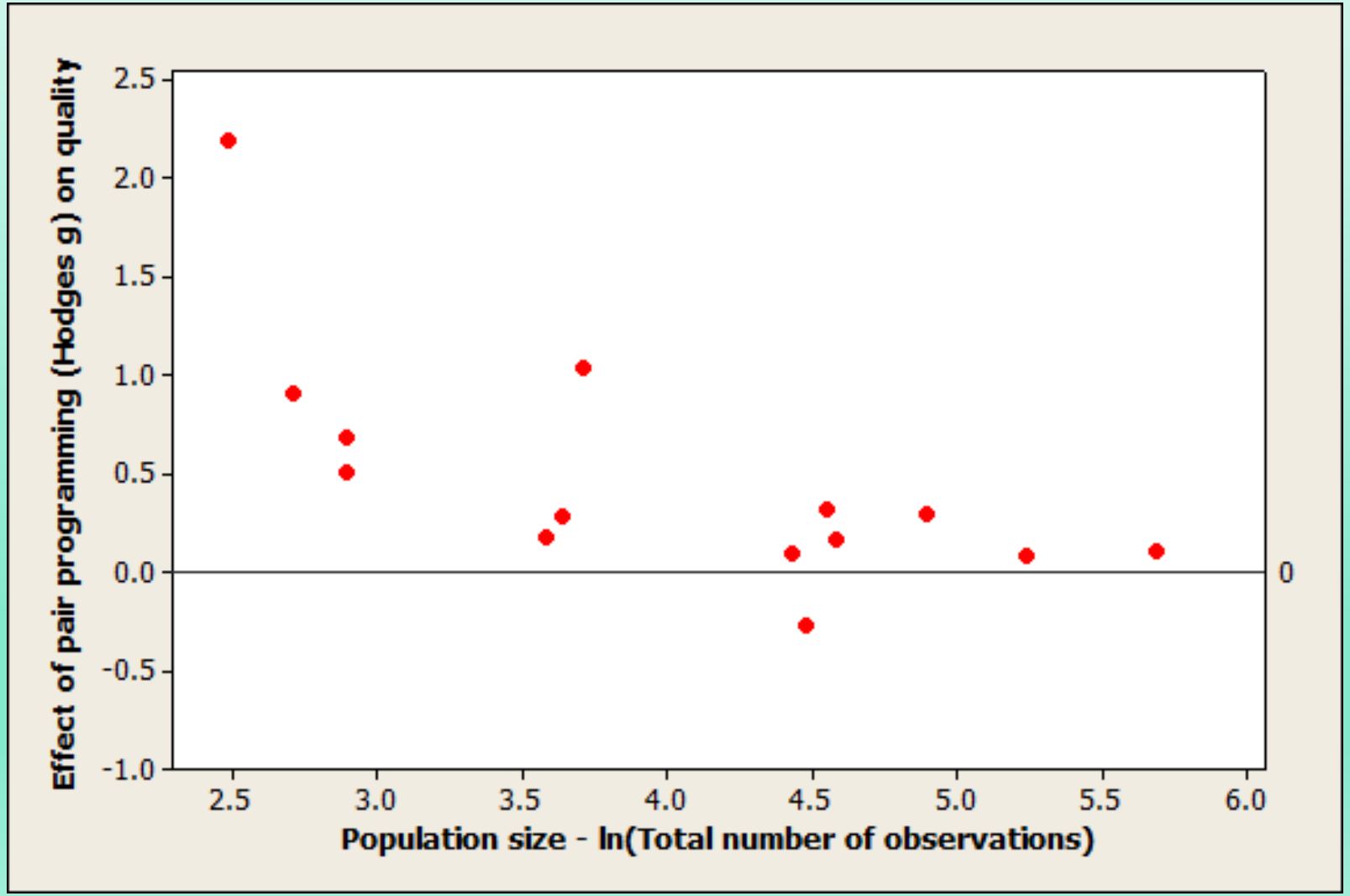

Hauser et al. (2014)[23], Nature
Gneezy et al. (2014)[22], Science
Janssen et al. (2010)[24], Science
Balafoutas and Sutter (2012)[18], Science
Pyc and Rawson (2010)[31], Science
Aviezer et al. (2012)[17], Science
Nishi et al. (2015)[30], Nature
Duncan et al. (2012)[20], Science
Karpicke and Blunt (2011)[25], Science
Derex et al. (2013)[19], Nature
Kovacs et al. (2010)[27], Science
Morewedge et al. (2010)[29], Science
Wilson et al. (2014)[36], Science
Rand et al. (2012)[33], Nature
Ramirez and Beilock (2011)[32], Science
Sparrow et al. (2011)[35], Science
Shah et al. (2012)[34], Science
Gervais and Norenzayan (2012)[21], Science
Kidd and Castano (2013)[26], Science
Lee and Schwarz (2010)[28], Science
Ackerman et al. (2010)[16], Science
Market belief (replicated, P < 0.05)
Market belief (not replicated, P > 0.05)
Survey belief (replicated, P < 0.05)
Survey belief (not replicated, P > 0.05)
Other researchers, when averaging their judgments, were amazingly good at guessing which results that would be possible to reproduce.
Studies with few observations, p-value close to 0.05, no convincing reason (theory) behind and news-friendly results were typically not reproduced.
0
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

This is further illustrated in: “Why most discovered true associations are inflated”, Ioannidis, Epidemiology, Vol 19, No 5, Sept 2008

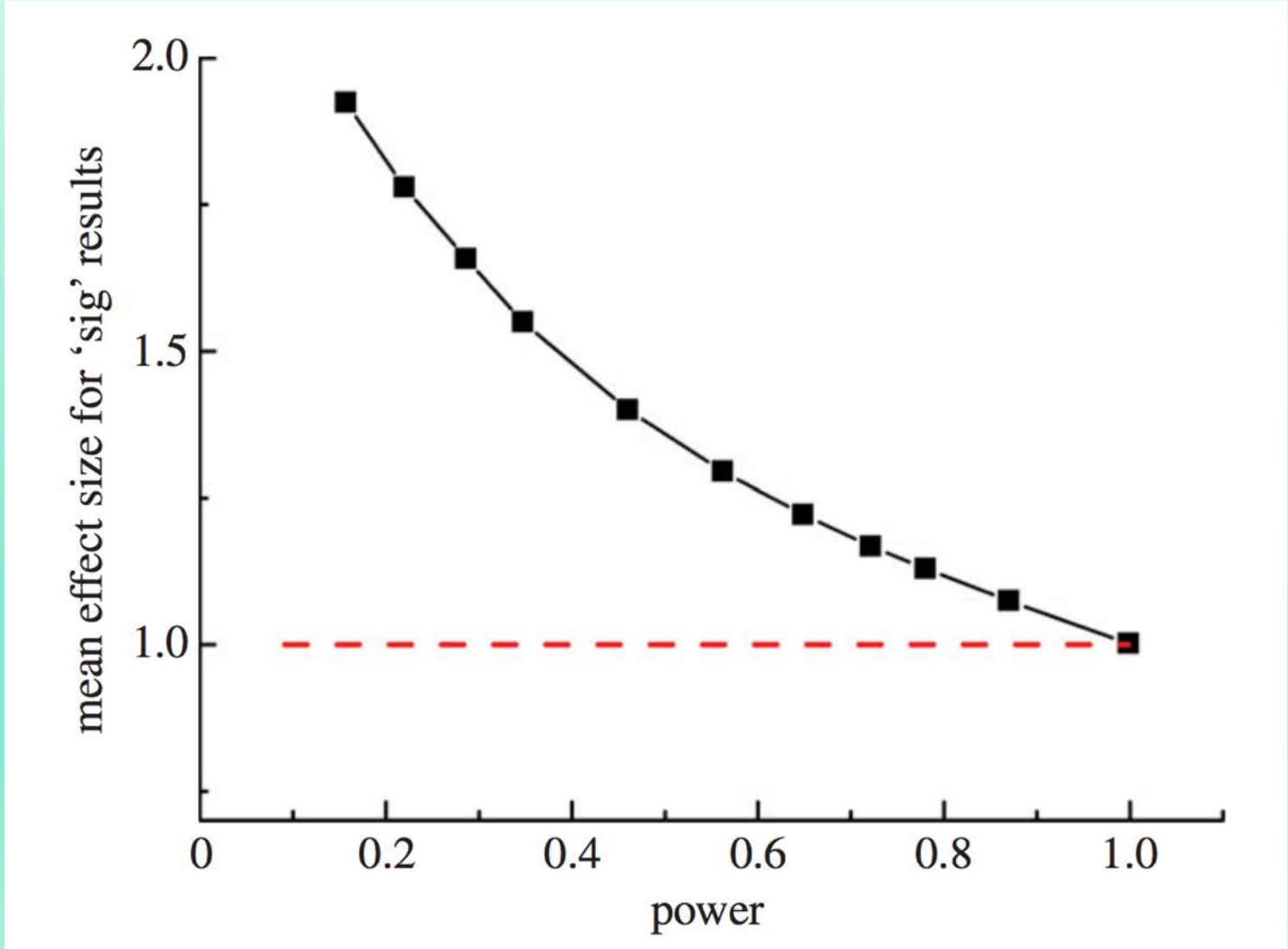


Example from empirical software engineering

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



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



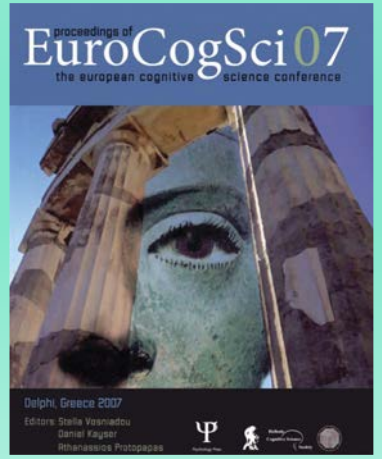
My own (in hindsight) obvious example of non-reproducible results

Context:

Lack of a good theory/reason to believe in the effect

Low statistical power (small scale)

Publication bias due to testing many characteristics and only publishing when finding "interesting" results



Jørgensen, M. (2017, September). Individual differences in how much people are affected by irrelevant and misleading information. In Proceedings of the European Cognitive Science Conference 2007. Taylor & Francis.

A short replication quiz

- **Original study**: Two groups with 20 developers in each. Difference, e.g., in productivity, between the two groups is d=0.7 (medium effect size), with p=0.049.
- **Question**: How much should you increase the sample size to have an 80% probability of finding p<0.05. Assume that the true effect size is 50% of that in the original study.
- **Answer:** Approx. 8 times more than in the original study!

We have in software engineering a large improvement potential when conducting and discussing replications:

- **Replication should NOT be about finding p<0.05 in the replicated study or not!!** (p-values will usually be all over the place)
- Replication studies should be about comparing **effect sizes**, **confidence intervals of effects sizes** and the use of **meta-analyses** to find the **aggregated effect sizes** and **intervals**.

Lessons learned: We need to …

Increase the statistical power (study size) of studies.

Reduce researcher and publications bias, e.g. through pre-registration of studies.

Stop thinking of replications as replicating p-values. Examine confidence intervals of effect sizes instead.

Another mistake with origin from psychology we should learn from …

PUNISHMENT VS. REWARD

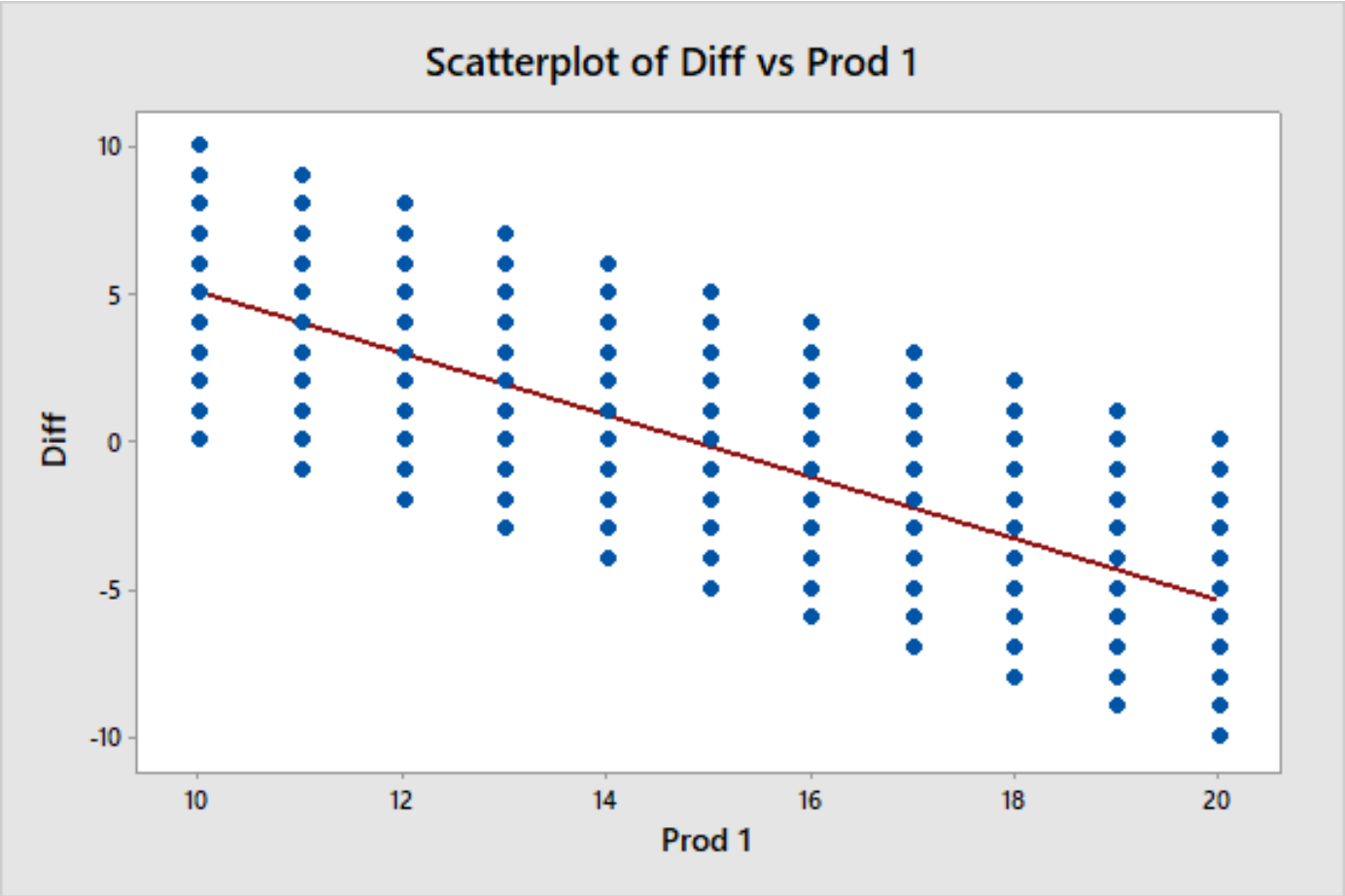
Coupling of variables ("Regression effects" and "Regression towards the mean")

EXAMPLE OF DIRECT COUPLING: Random productivity values

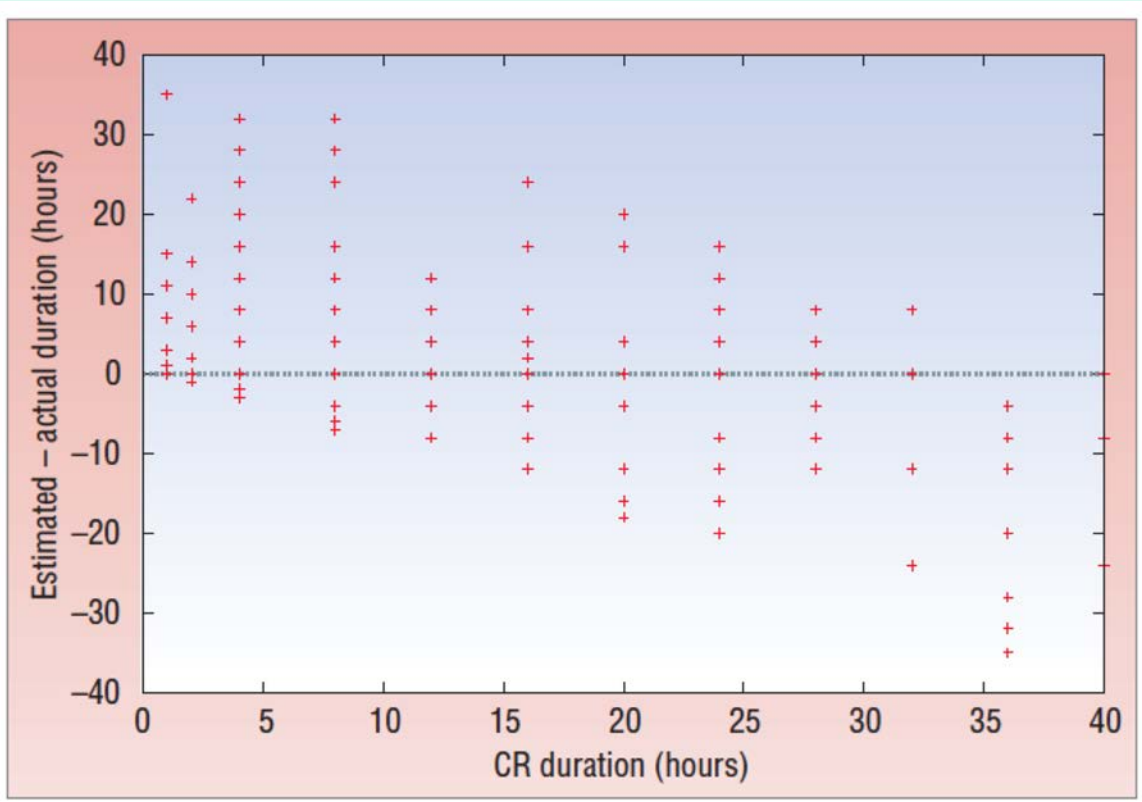
$Prod_1$ = randbetween(10,20) = initial productivity

$Prod_2$ = randbetween(10,20) = new productivity

Diff = Prod_increase = $Prod_2 - Prod_1$



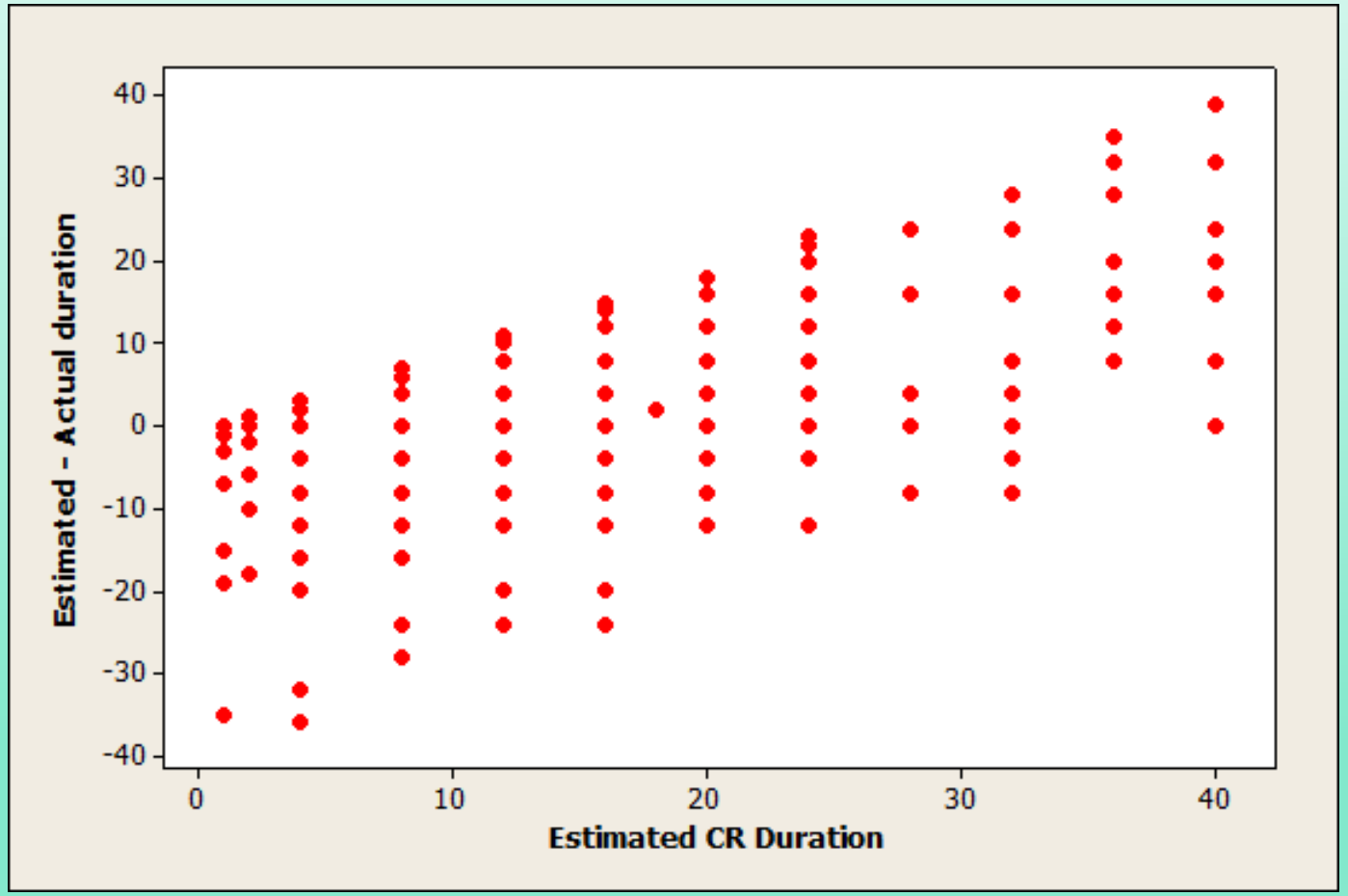
Graph from a software engineering paper

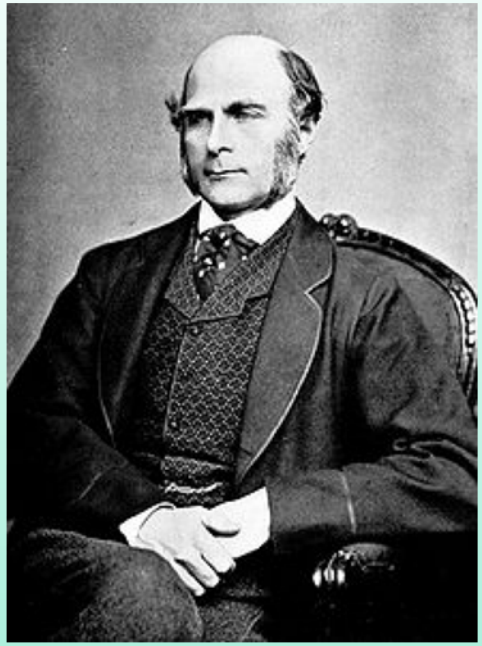


X-axis: Actual hours
Y-axis: Estimated hours – actual hours
Correlation: Around -0.7
Interpretation: More estimation over-optimism on larger tasks

Same data, now with the estimated (before it was the actual) effort as the x-axis

Opposite interpretation: Less over-optimism on larger tasks

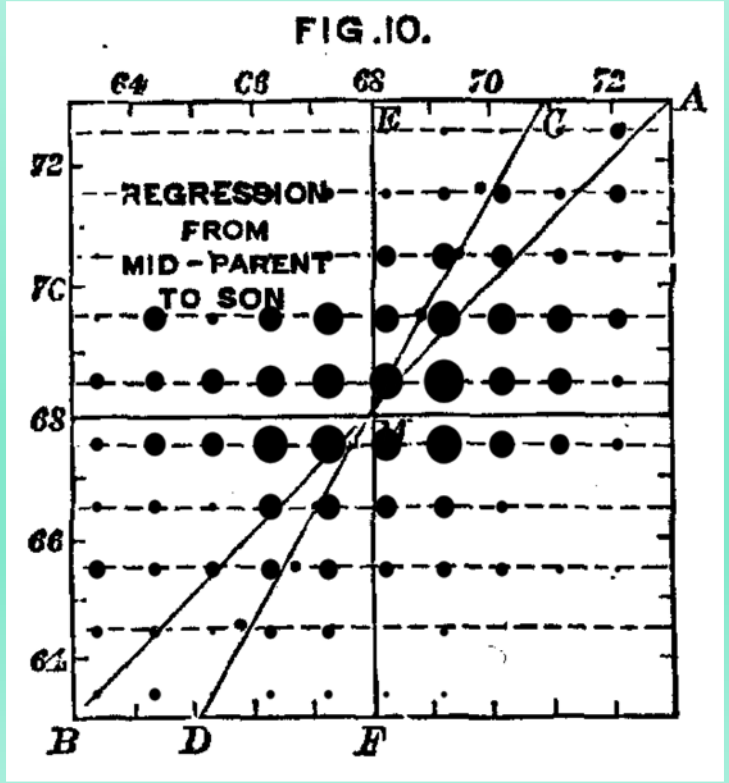




Example of coupling through common random variation element (Regression towards the mean)

Sir Francis Galton ("*Filial regression to mediocrity*", 1886):

- The inventor of regression analysis - and the first to be fooled by it.
- Finding: Children of tall parents were on average lower than their parents.
- But, parents of tall children were, at the same time, on average shorter than their children!



Lessons learned: Know your analysis methods well and avoiding reporting statistical artefacts

- Avoid shared component (mathematical coupling) between the explained and the explaining variables.
- Common situations reporting statistical artefacts (X and Y are stochastic variables):
  - Analyse he relation between (Y-X) and X, e.g., analyse the improvement (prod2-prod1) relative to starting position (prod1) after introducing a new tool.
  - Compare the category means of Y/X, where X is used in the categorization, e.g., analyse the mean productivity (Size/Effort) of small, medium and large projects where Size is used to categorize the projects.
  - Find the (OLS-based) regression slope (b) of $Y = a + bX$, e.g., analyse whether there is an economy-of-scale, i.e., $b<1$, using regression analysis on $Effort = a + bSize$)
    [Here the, more hidden, shared component it caused by that the random error of Y includes the random error of X and deflates the b-value.]

ELSEVIER

Journal of Systems and Software

Volume 85, Issue 11, November 2012, Pages 2494-2503

Interpretation problems related to the use of regression models to decide on economy of scale in software development

Magne Jørgensen [a, b] ✉, Barbara Kitchenham [c]

ELSEVIER

International Journal of Project Management

Volume 30, Issue 7, October 2012, Pages 839-849

How does project size affect cost estimation error? Statistical artifacts and methodological challenges

Magne Jørgensen [a, b] ✉, Torleif Halkjelsvik [b] ✉, Barbara Kitchenham [c] ✉

Most of our research would not have been accepted in empirical psychology simple because of being too shallow.

We should aim for **deeper understanding** of core mechanisms

We should aim for/use theories, models, explanations – not just study effects.
This requires studies **designed to understand** not only measure effects.

Example from psychology: Prospect theory

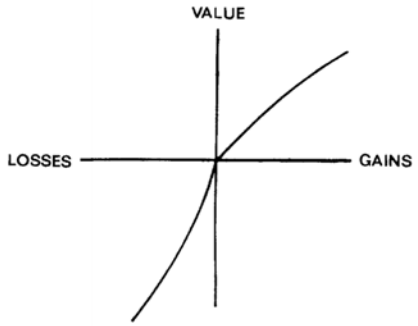


FIGURE 3.—A hypothetical value function.

- 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 ;-)

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.

Other things psychology do differently and we should learn from

- Systematic Literature Reviews are written by the senior researchers with long experience in a research field. It is NOT acceptable that the least knowledgeable in a field summarize the knowledge.
- Conferences there mainly to share ideas and present on-going work (abstract only), journals for publishing results.
- Pre-registration of studies (avoiding publication bias and post-hoc analyses) and up-front power analysis (avoiding low power studies) more and more common.
- More focus on testing/creation of theories, transparency and rigor in qualitative studies (as far as I can see, see e.g., APA Reporting standards)

Other changes I think we would benefit from

- Software engineering students should, similarly to psychology students, be introduced to psychology and empirical methods as part of their curriculum
- Software engineering researchers should spend more time on finding, reading and using relevant psychology research results
- Software engineering researchers should collaborate more with researchers in psychology (e.g., within behavioural economics, sociology, …)
  - My experience is that they will find us interesting.

The end is here …