Evidence-Based Methods in Cyber Security Research

Thomas Gross

UK Cyber Security PhD Winter School 2020, 15-01-2020

Based on research in the
UK Research Institute in the Science of Cyber Security (RISCS)

Newcastle University

Collaborators & Funders

- Angela Sasse (Ruhr-University Bochum)
- Audrey Linden (Northumbria)
- Johannes Honekopp (Northumbria)
- Kovila Coopamootoo (Newcastle)
- Malte Elson (Ruhr-University Bochum)
- Roy Maxion (CMU/Newcastle)
- Uchechi Nwadike (Newcastle)

Aim & Scope

- **Pragmatic 101 & Sensitization:**
  How to use evidence-based methods?

- **State-of-Play:**
  How are we doing in EBM in cyber security?
  Contrasting ideal states of evidence-based experimentation with empirical results.

- **Scope:**
  Quantitative, experiments, user studies.

tl;dr

- Evidence-based methods have great potential, **but** our field's implementation is often flawed.
- Design for falsification & reproducibility is key.
- Empirical analysis of state-of-play highlights issues in individual studies as well as the field.
- Imprudent design choices and biases cause a low posterior probability of positive reports.
- Replications crucial, yet virtually non-existent.

SETTING THE STAGE

What are evidence-based methods?

- Empirical methods for attaining knowledge.
- **Part of: Evidence-based practice (EBP**), any practice that relies on scientific evidence for guidance and decision-making.
- **Employ: Scientific method**, mathematical and experimental technique employed in the sciences. More specifically, it is the technique used in the construction and testing of a scientific hypothesis.

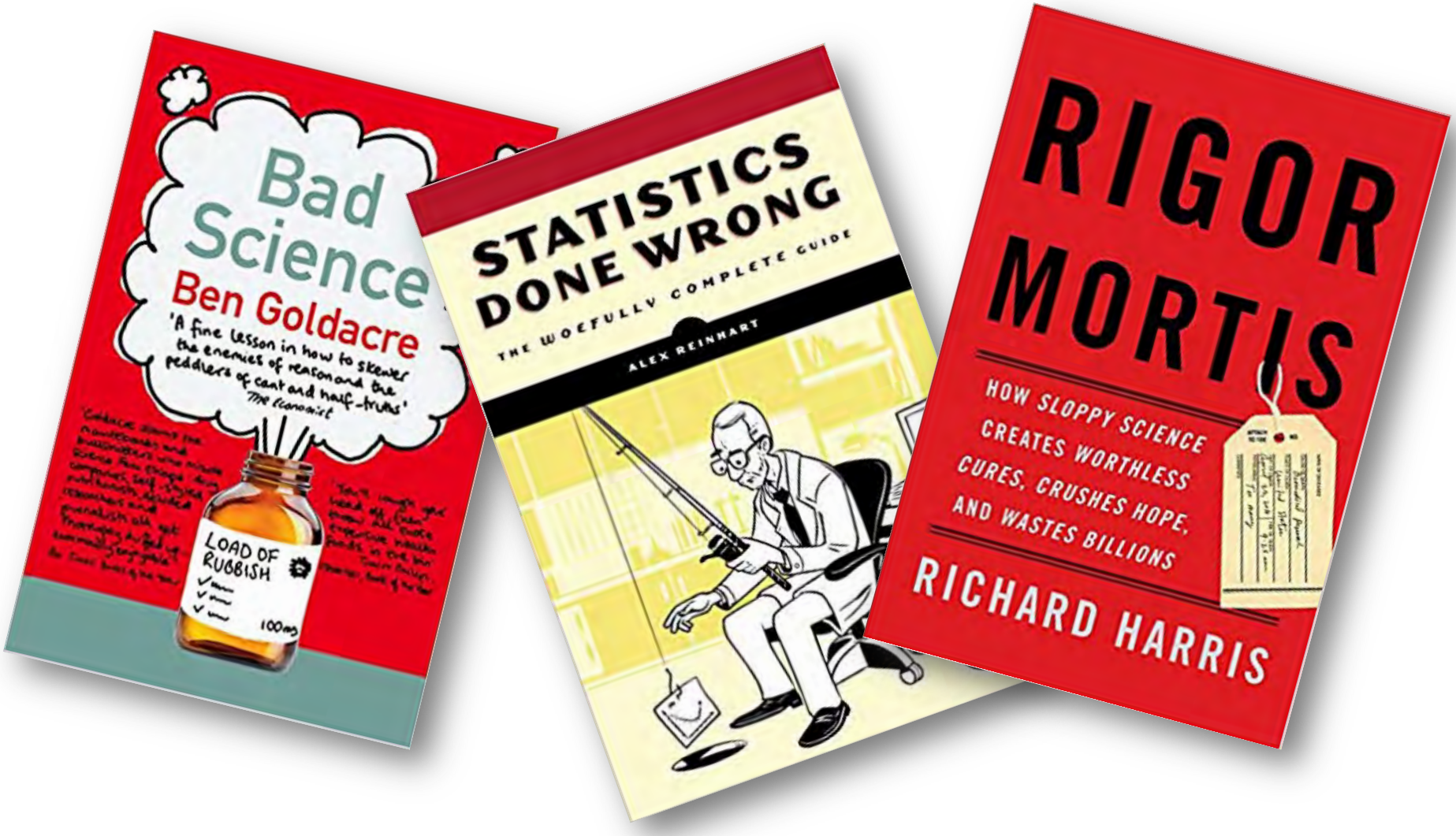
-- Encyclopædia Britannica

The Promise of *Evidence-Based Methods*

- “Science of Security”
- Scientific, evidence-based methods to be the bedrock of the foundations.
- Evidence to support our every claim.

Bad Science Under Fire

Some Light, Outrageous, and Disturbing Bedside Reading



Why Most Published Research Findings Are False

- Ioannidis analyzed *posterior* likelihoods of research findings
- *Positive Predictive Value*
- Considers bias and *prior* likelihood estimates.
- Showed that most published positive findings must be *false*.

Essay

Open access, freely available online

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

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

[Ioannidis, 2005 – Why most published research findings are false]

Begley's Bombshell
Replication Failure in Cancer Research

- Begley et al. attempted to reproduce **53** landmark studies in a 10 years timeline.
- Only **6** studies (**11%**) could be reproduced.
- Non-reproduced studies still highly cited.


COMMENT

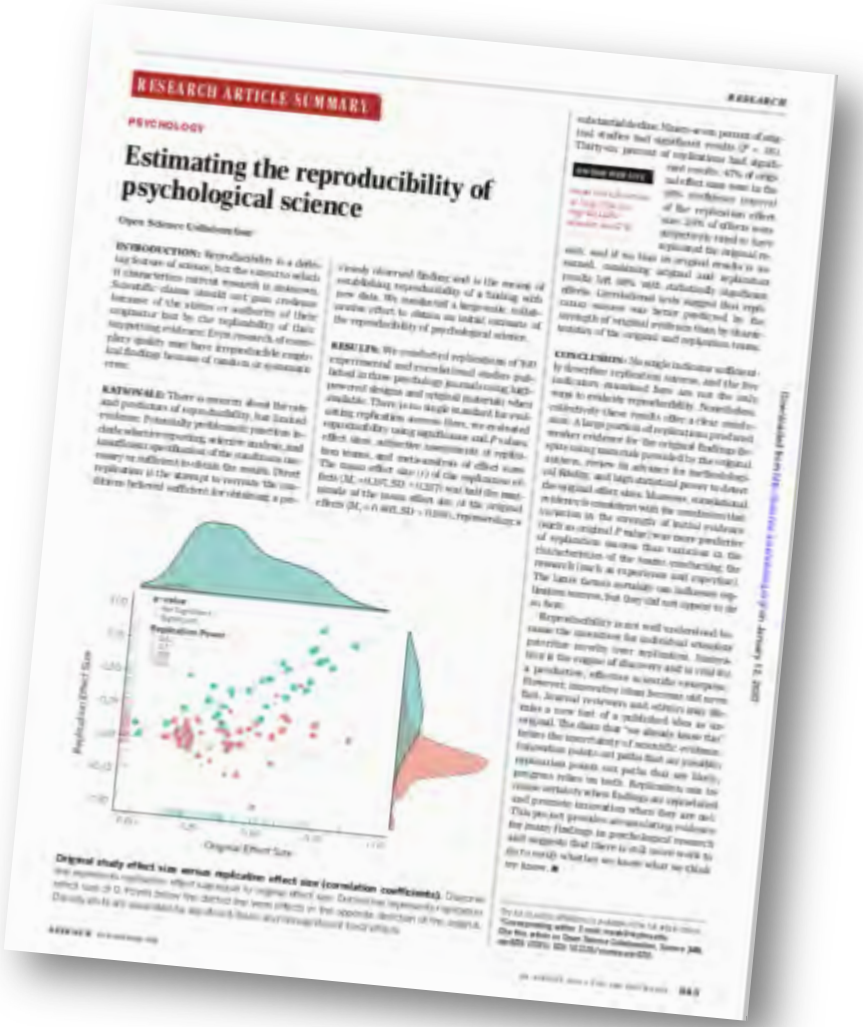
Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

[Begley and Ellis, 2012 – Raise standards for preclinincal cancer research; cf. Harris, Rigor Mortis]

Replication Crisis in Psychology

- Open Science Collaboration attempted replications of **100** psychology studies.
- Replications had **half** the mean effect size of original studies.
- **Only one third** of replications had significant results ($p < .05$)



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration

[Open Science Collaboration, 2015 – Estimating the reproducibility of psychological science]

The Promise of *Evidence-Based Methods*

- “Science of Security”
- Scientific, evidence-based methods to be the bedrock of the foundations.
- Evidence to support our every claim.

SCIENTIFIC METHOD

Steps of the Scientific Method

A Pragmatic View

- Intriguing observation
- Formulation of a research question
- Formulation of explanatory theory/hypothesis
- Creating experiment to test of hypothesis
- Devise ingenious analysis of the data
- Attempt to prove theory/hypothesis

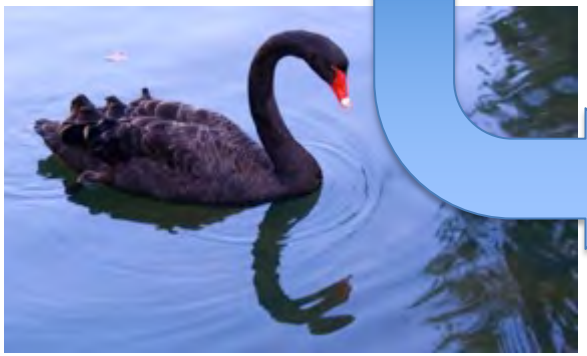
wrong!

Falsifiable Hypotheses and Theories

The Problem of Induction

Even many observations of *white swans* **cannot verify** the theory.

Theory: "All Swans are white"



One observation of a *black swan* **falsifies** the theory.

Falsifiability of a hypothesis or theory is the inherent possibility that it can be proved false.

[Popper, 1959; Carroll et al., 2012]

[Slide from tutorial on evidence-based methods..., IFIP Summerschool, w/ Coopamootoo]

Reproducibility

- **Roots:** transparency, not corroboration
- **Methods reproducibility**, provision of enough detail about study procedures and data so the same procedures could, in theory or actuality, be exactly repeated.
- **Results reproducibility** (also **replicability**), obtaining the same results from conduct of an independent study whose procedures are closely matched to the original study.

[Goodman et al., 2016 – What does research reproducibility mean?]

Validity

- **Validity,** trust or correctness, a correspondence between a proposition how things work in the world and how they really work.
- **Internal Validity,** extent to which a causal conclusion is warranted by a study.
- **External Validity,** extent which a conclusion of a study can be applied outside of the context of the study itself (generalized across situations, people, stimuli and times).

[Maxwell and Delaney 2004, Designing Experiments and Analyzing Data]

Experiment

An investigation in which the investigators have sufficient control of the system under study, in particular to be able to determine the assignment of different units of study to different conditions.

[Oxford Dictionary of Statistical terms, 2004]

A Conversation with Nature

- “We might think of an experiment as a conversation with nature, where we ask a question and listen for an answer.”

  -- Martin Schwarz

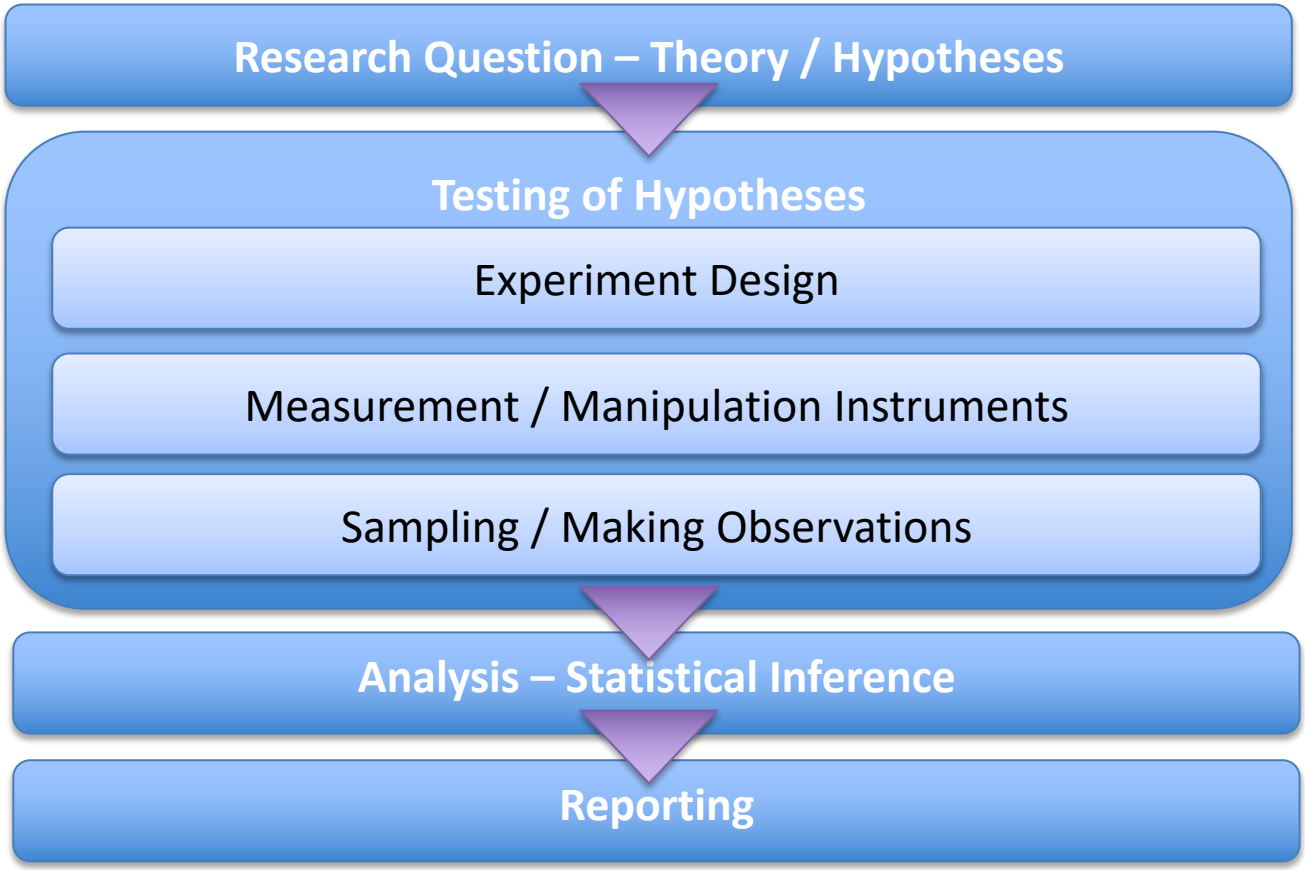
- ”There is in fact no such thing as direct observation… Every scientific observation is filtered through a instrument of some sort.”

  -- Steven Goodman

- Manipulation & measurement rely on **instruments**.

Steps of the Scientific Method

A Pragmatic View



Statistical Hypothesis Testing

Research Question

Relevant?
Clear?
Focused?
Concise?

Hypotheses

Null H_0

Alt. H_1

Testable?
Falsifiable?

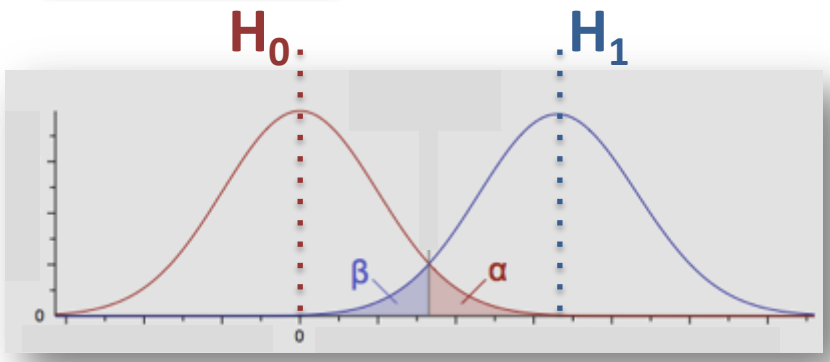
Stats Setup

Assumptions?
Test Statistic?
Distribution?
Sig. Level α

Observation

Statistical Inference

Evaluate test statistic & critical region.
Reject H_0/H_1



 [Lehmann and Romano (2005) Testing Statistical Hypotheses]

Statistical Inference Decision Matrix

OUR CONCLUSION \ REALITY	H_0 is TRUE H_1 is FALSE In reality: no difference	H_0 is FALSE H_1 is TRUE In reality: there is a difference or a gain.
We accept H_0. We reject H_1. We say: "There is no difference, no gain."	**CONFIDENCE LEVEL** **1-α** (Odds of saying that there is no difference, when in fact there is none)	**TYPE II ERROR** **β** (Odds of saying that there is no difference, when in fact there is one)
We reject H_0. We accept H_1. We say: "There is a difference, a gain."	**TYPE I ERROR** **α** (Odds of saying that there is a difference, when in fact there is none)	**POWER** **1-β** (Odds of saying that there is a difference, when in fact there is one)

[Adapted from the Social Research Methods Knowledge Base: Statistical Power]

p-Value vs. Effect Size

Significance

Likelihood of observing a result equal or more extreme as the sample, assuming H_0.

$Pr(Observation \mid H_0) \neq Pr(H_0 \mid Observation)$

→ ***p*-value**

Importance

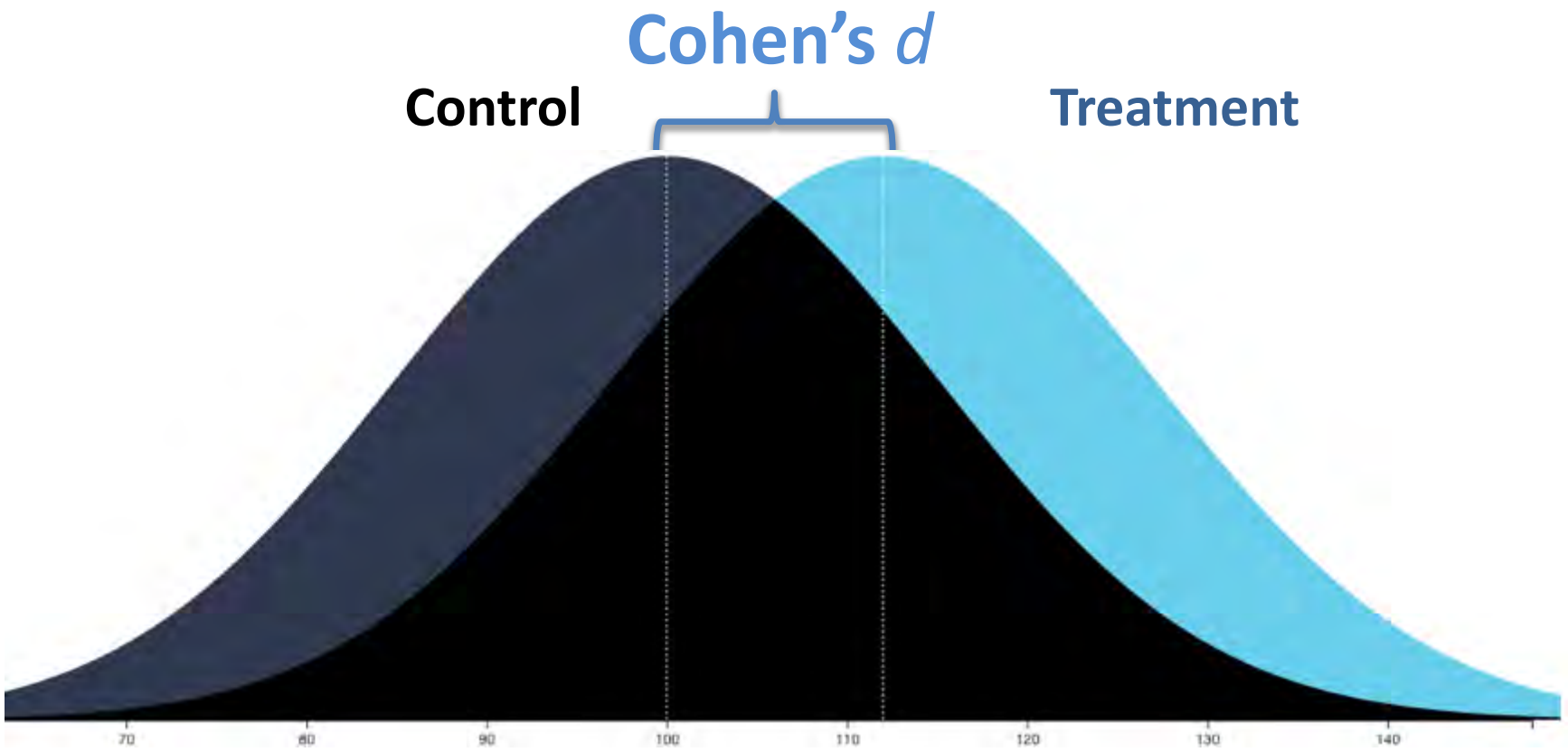
Standardized magnitude of a phenomenon.

Determines whether the effect is non-trivial and relevant for practice.

→ **Effect Size**

Effect Size

Online

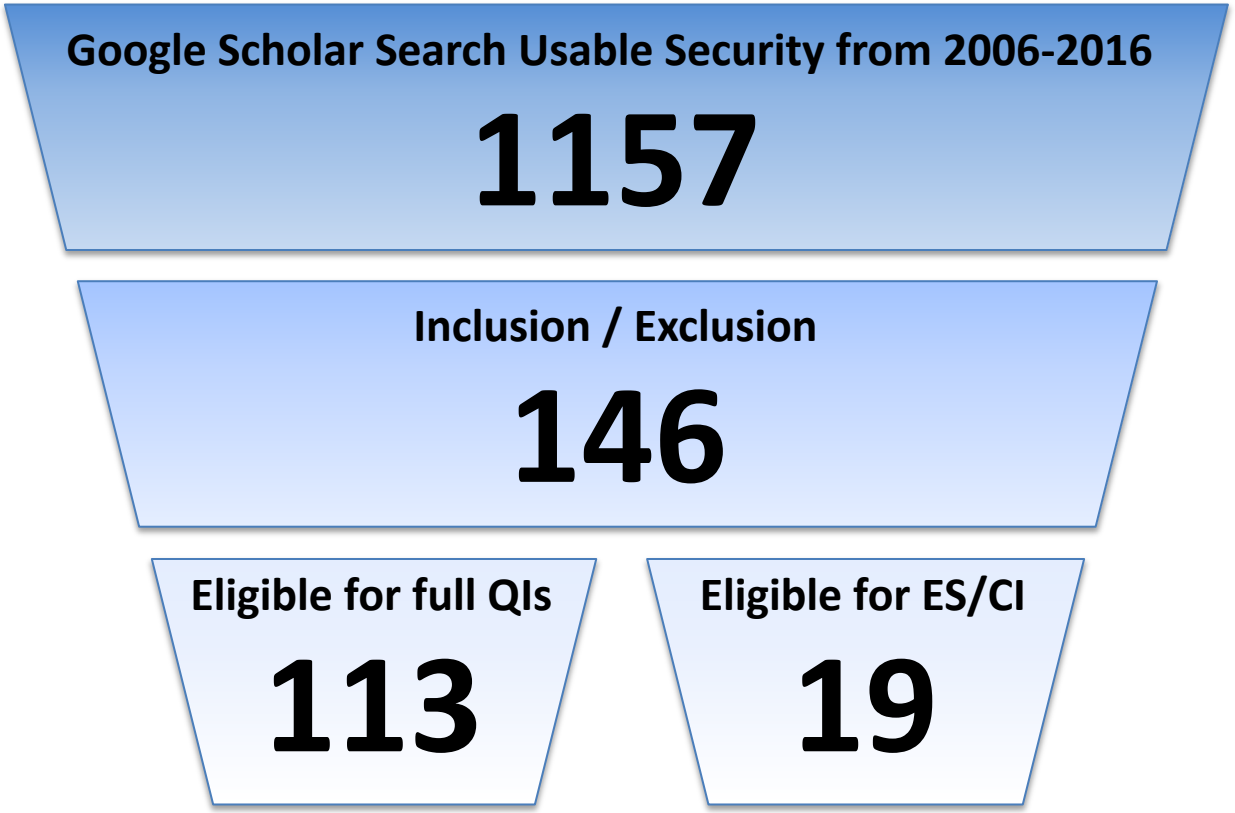


Cohen's *d*: Difference between Control and Treatment standardized by the common within-population standard deviation σ.

[Graph by Kristoffer Magnusson, https://rpsychologist.com/d3/cohend/]

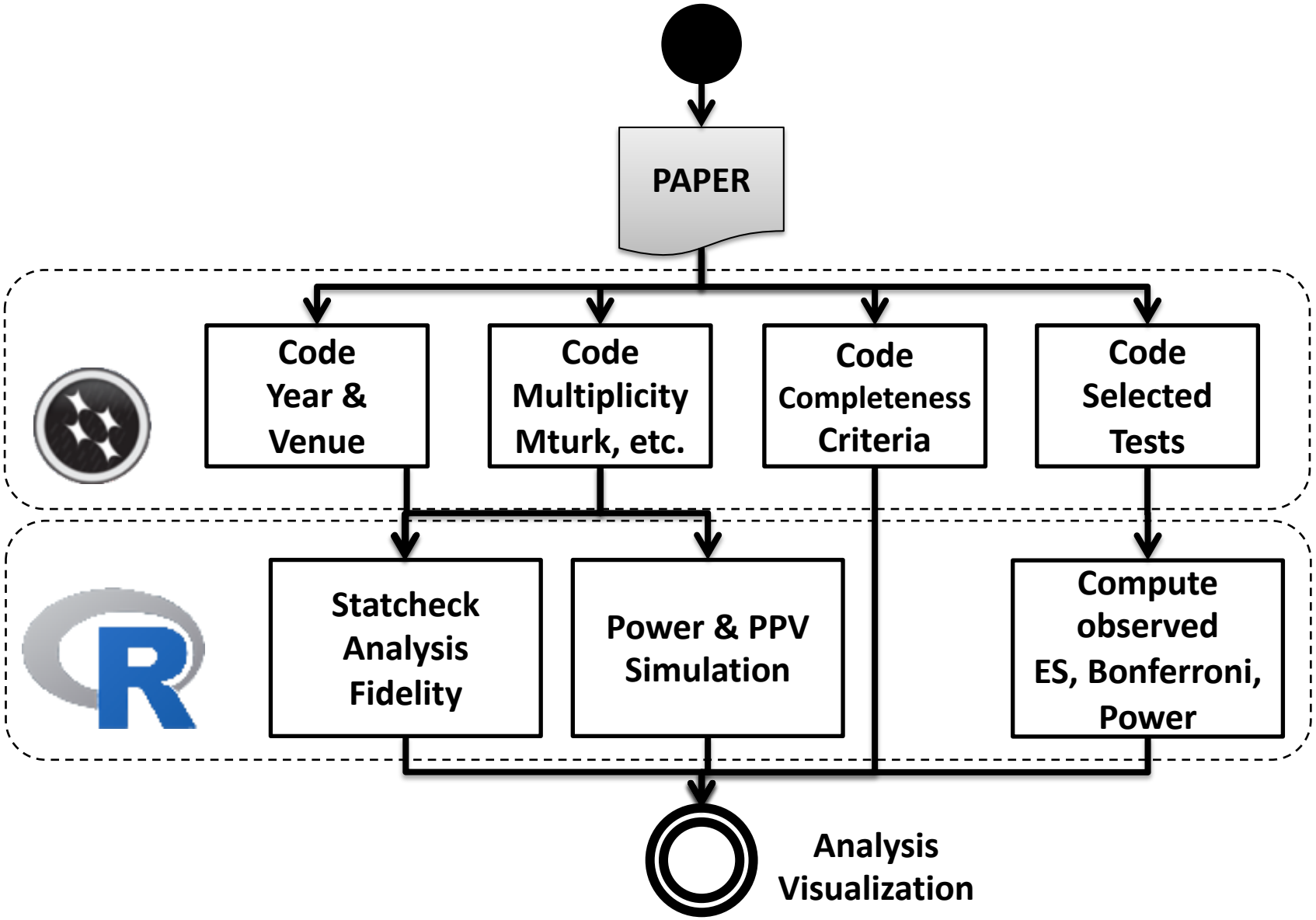
SYSTEMATIC REVIEW OF STATE-OF-PLAY

RISCS-Funded SLR 2016/17:
Evaluation of Experiments on Human Factors in Security and Privacy



[Joint work with Kovila Coopamootoo, RISCS, 2016/17]

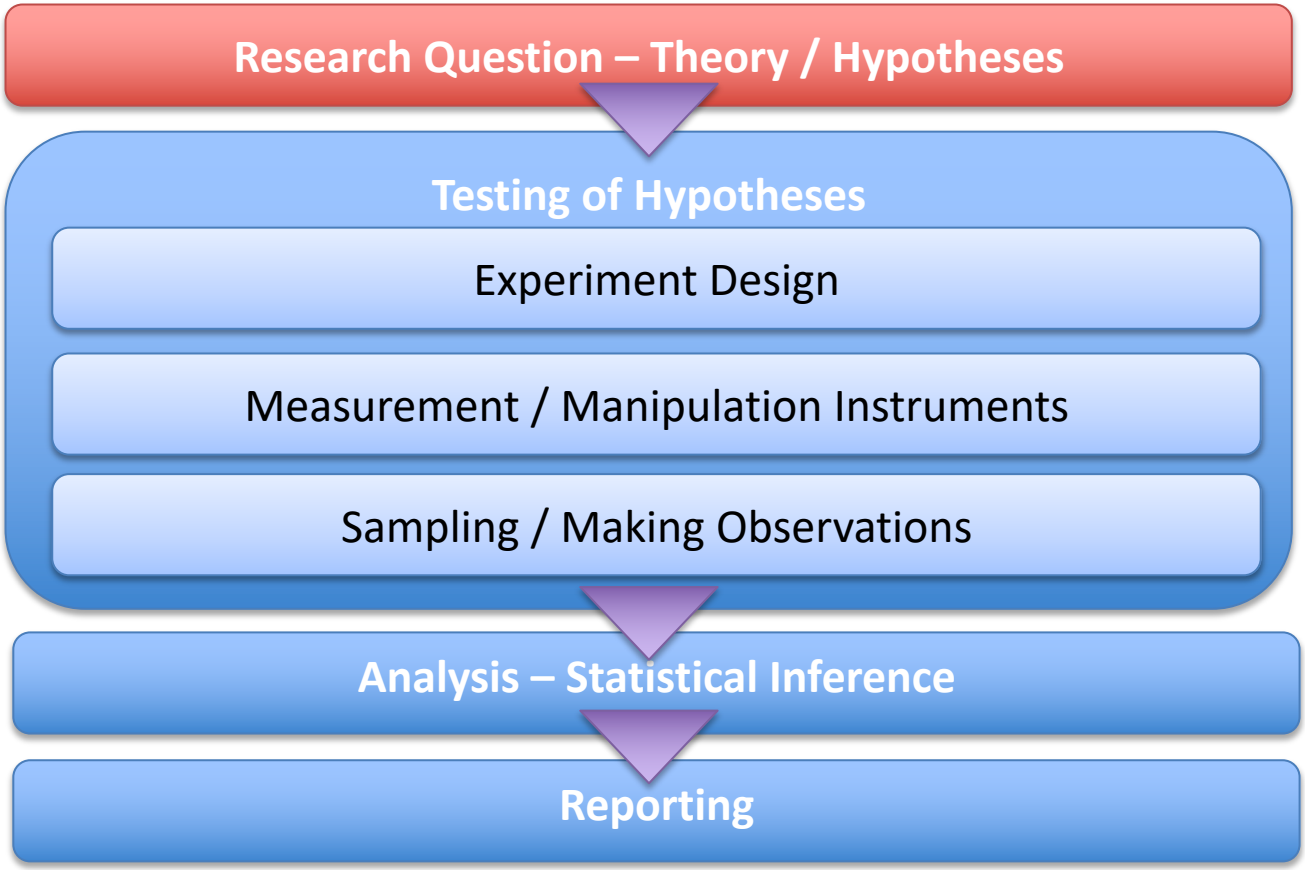
Workflow of SLR-Related Studies



STATE-OF-PLAY OF INDIVIDUAL STUDIES

WITHOUT A BEARING

Steps of the Scientific Method



Research Intent

"It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude."

Sir Harold Jeffreys

- Science is a lot about asking the right questions.
- Framing research questions and operationalizing them into testable hypotheses.

[Jeffreys H. *Theory of Probability*, 1939. Quoted from Etz, Vandekerckhove. A Bayesian Perspective on the Reproducibility Project, 2016.]

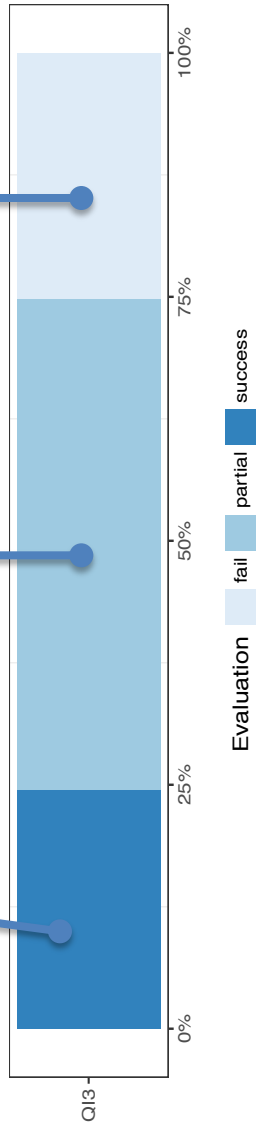
What if the RQ/hypotheses are ill-defined?

- Muddle the waters.
- Inhibit the sound interpretation of the results.
- Perpetuate researcher bias.
- Robs statistical tests of their foundations.

In security and privacy user studies, **26%** did not state the research question.

Of a sample of **146** studies

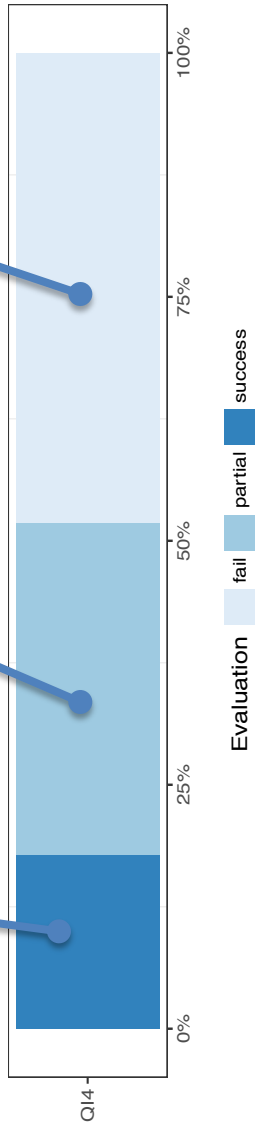
- One quarter (**26%**) failed to state research question and hypotheses at all.
- Half (**50%**) left the research questions and their operationalization under-specified.
- One quarter (**24%**) offered sound research questions and hypotheses.



In security and privacy user studies, only **52%** reported limitations somewhat.

Of a sample of **146** studies

- Half (**48%**) failed to state limitations of design or execution at all.
- One third (**34%**) reported some limitations, biases and confounders.
- Less than one fifth (**18%**) offered a complete analysis of limitations.



Research Intent

Dos

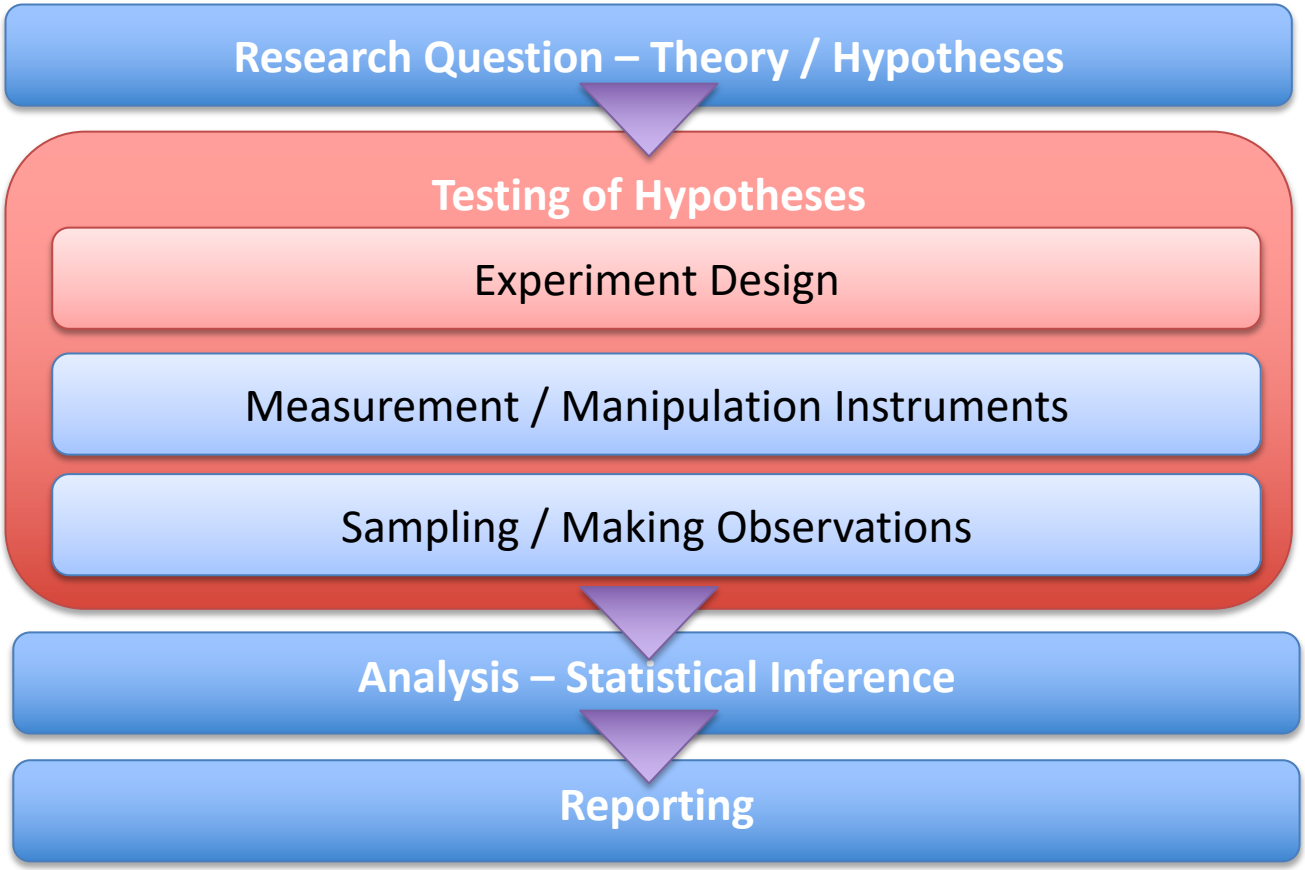
- Specify clear, focused, concise research questions.
- Develop meaningful statistical hypotheses *a priori* of data collection.
- Develop a sound operationalization.
- Write a structured abstract incl. method and intended outcome.
- Pre-register the study.

Don’ts

- Do not start the study with muddled research intent.
- Do not switch track in terms of hypotheses once the study is underway.
- Do not omit or change the research question / hypotheses / operationalization for reporting.
- Do not omit limitations of your study.

ROADS TAKEN

Steps of the Scientific Method



Sound Design

“It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.“

– Richard P. Feynman

“The first principle is that you must not fool yourself and you are the easiest person to fool.”

– Richard P. Feynman

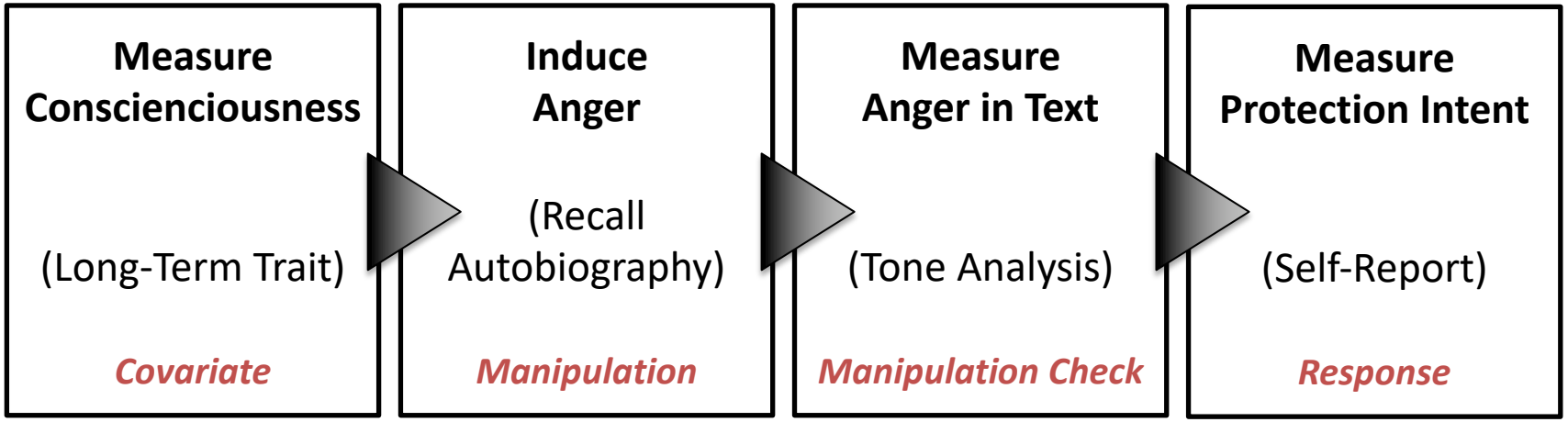
an experiment design

There are two ways of constructing a soft[illegible]sign; one way is to make it so simple that there are obviously no deficiencies, and the other way is to make it so complicated that there are no obvious deficiencies. The first method is far more difficult.

Cheekily adapted from Tony Hoare

What's wrong with this picture?

Adapted from a procedure of a published study, variables substituted.

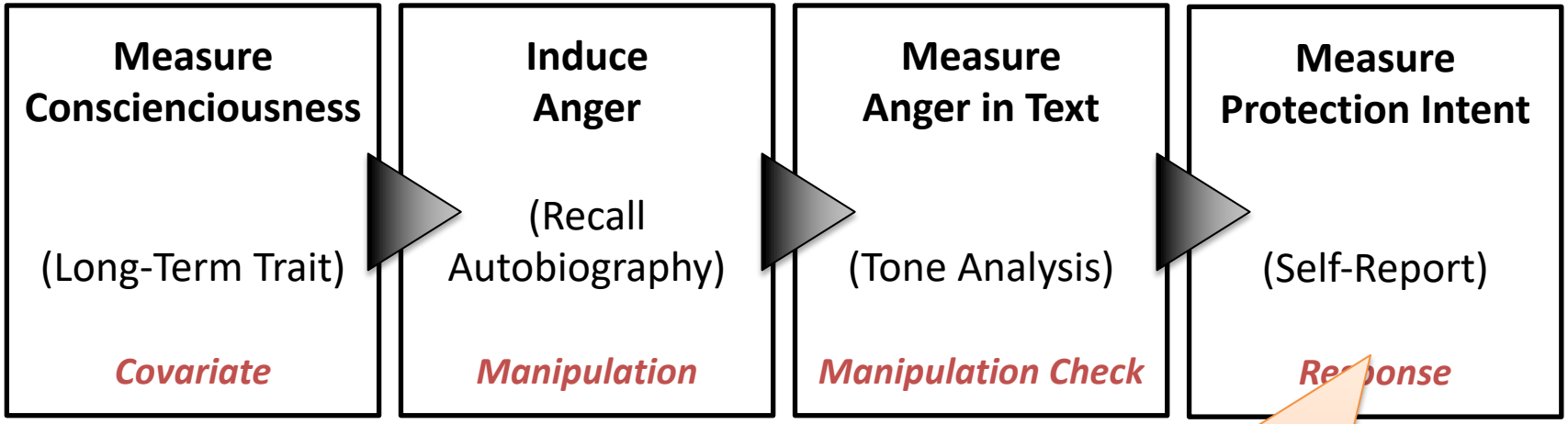


Hypotheses:
Induced anger impacts protection behavior.

Claim made:
- Change in anger significantly impacts change in conscientiousness.
- Conscientiousness significantly impacts protection behavior.

What's wrong with this picture?

Adapted from a procedure of a published study, variables exchanged.



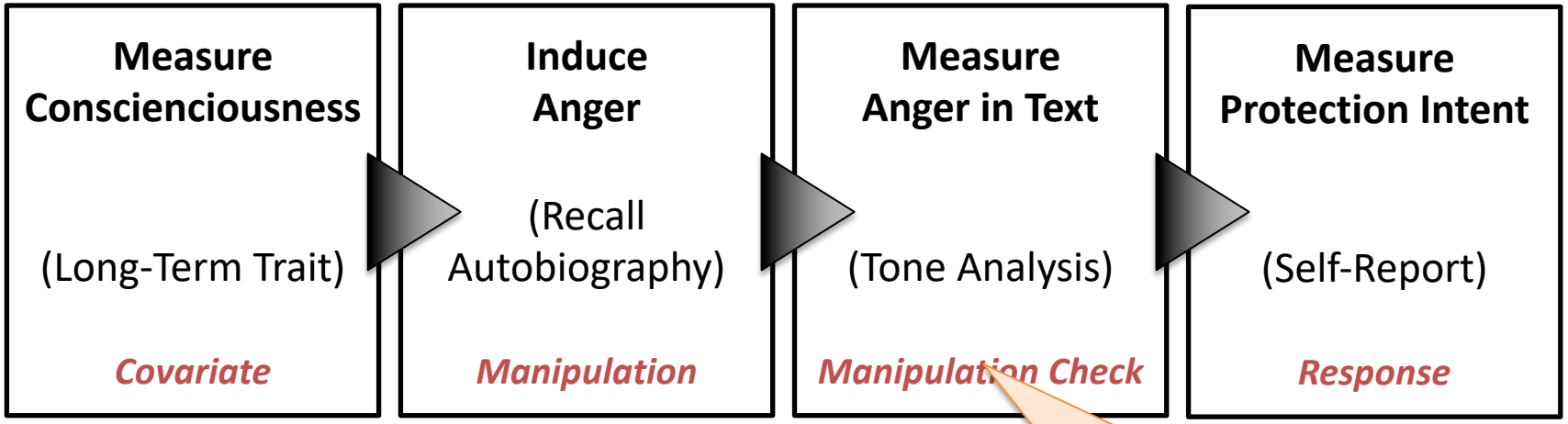
Hypotheses:
Induced anger impacts protection behavior.

Claim made:
- Change in anger significantly impacts cha
- Conscientiousness significantly impacts p

Unreliable Response:
Self-reported intent does not necessarily translate to behavior change.
Impacted by social desirability.

What's wrong with this picture?

Adapted from a procedure of a published study, variables exchanged.



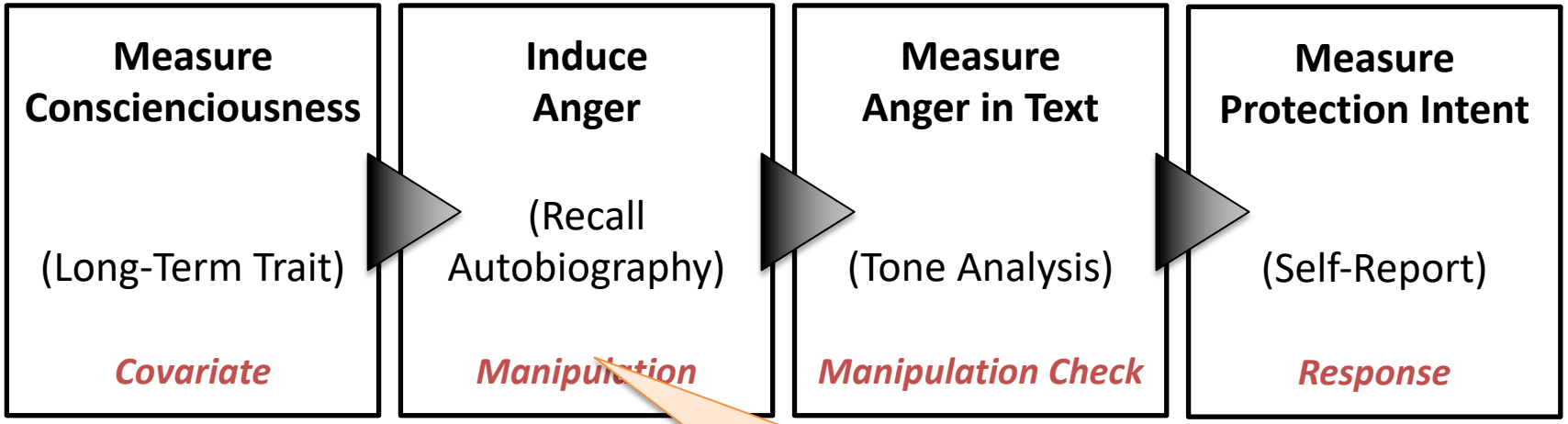
Hypotheses:
Induced anger impacts protection behavior.

Claim made:
- Change in anger significantly impacts change in c
- Conscientiousness significantly impacts protectio

Instrument Ill-Suited:
Anger tone present in text written by participant does not measure the participant's actual affective state.

What's wrong with this picture?

Adapted from a procedure of a published study, variables exchanged.



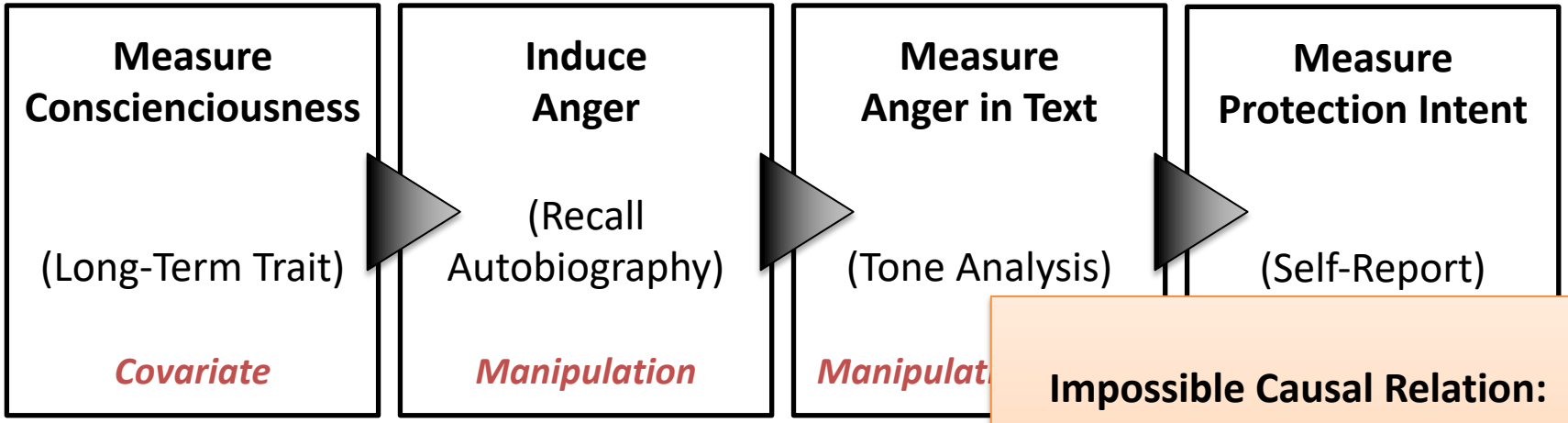
Hypotheses:
Induced anger impacts protectio

Claim made:
- Change in anger significantly i ss.
- Conscientiousness significantl

Valid Intervention?
How effective is autobiographical recall in changing participant's state?
Evidence for validation?
Effect sizes expected?

What's wrong with this picture?

Adapted from a procedure of a published study, variables exchanged.



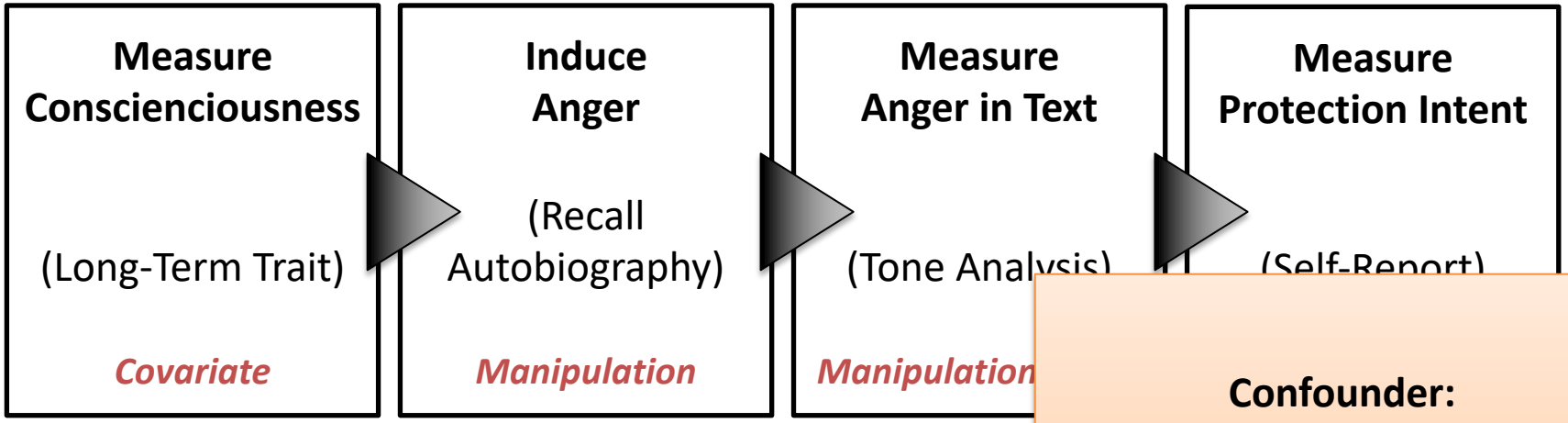
Impossible Causal Relation:
(Trait) conscientiousness measured before manipulation. Manipulation could not have affected the trait measure.

Hypotheses:
Induced anger impacts protection behavior.

Claim made:
- Change in anger significantly impacts change in conscientiousness.
- Conscientiousness significantly impacts protection behavior.

What's wrong with this picture?

Adapted from a procedure of a published study, variables exchanged.



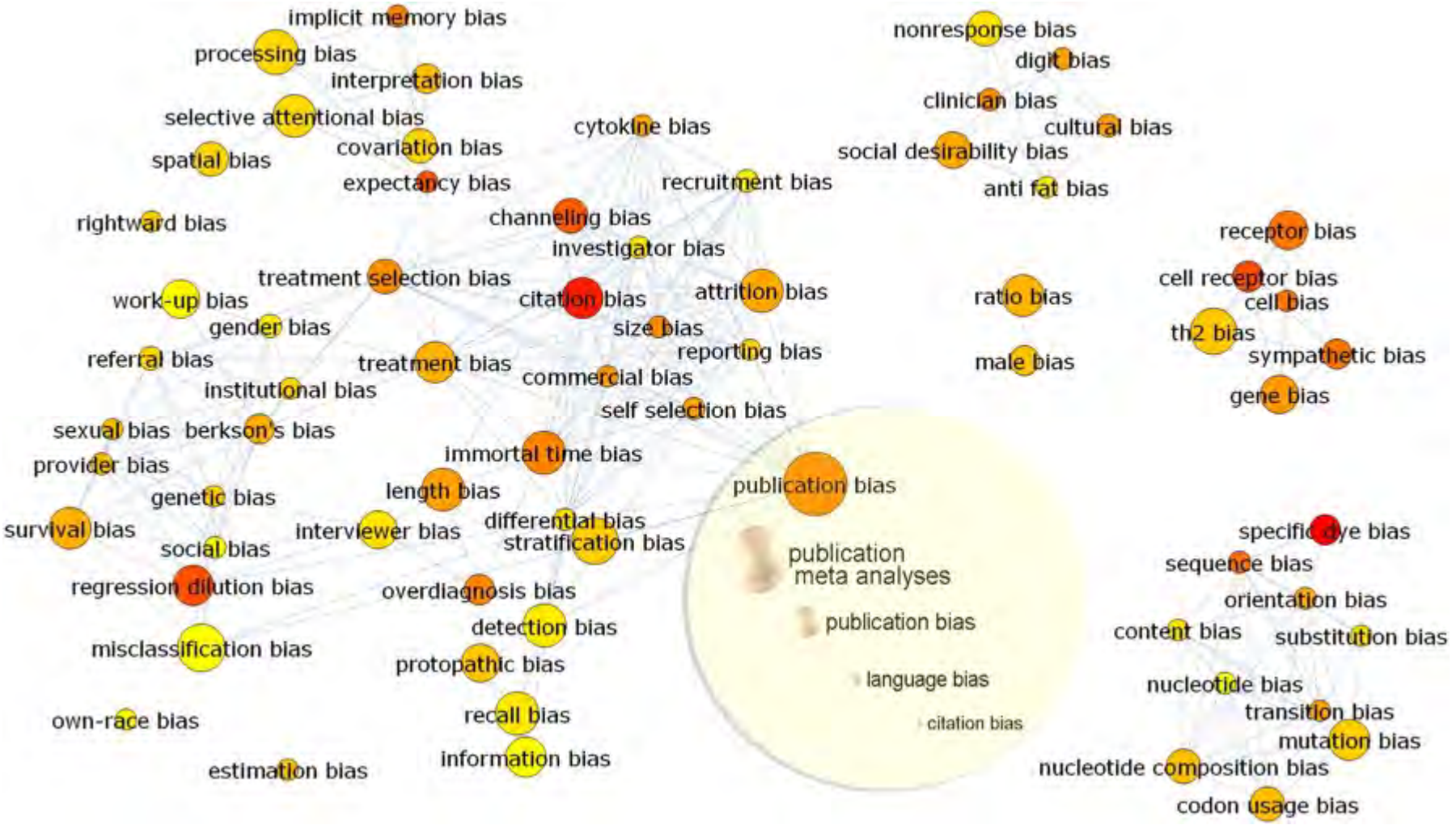
Confounder:
Consciousness could impact diligence in autobiographical recall and protection intention

Hypotheses:
Induced anger impacts protection behavior.

Claim made:

- Change in anger significantly impacts change in conscientiousness.
- Conscientiousness significantly impacts protection behavior.

Biases



[Chavalarias and Ioannidis 2010, Science mapping analysis characterizes 235 biases in biomedical research]

Sound Design

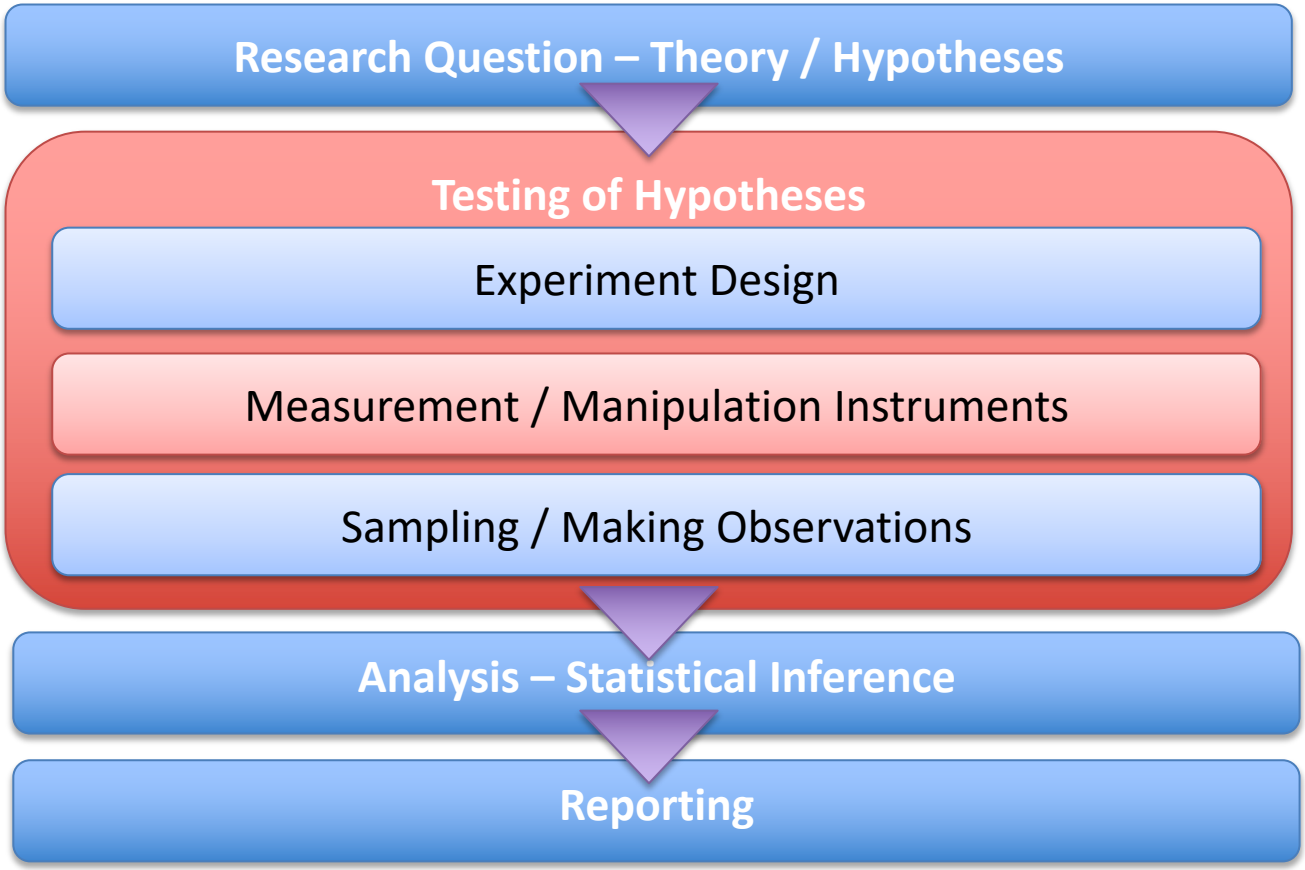
Dos

- Embrace clarity and simplicity.
- Stay true to course: Prove your theory *wrong*.
- Measure twice, cut once.
- Get a critical friend early on.
- Use standard designs for given purpose.
- Endeavour to minimize biases wherever you can.

Don’ts

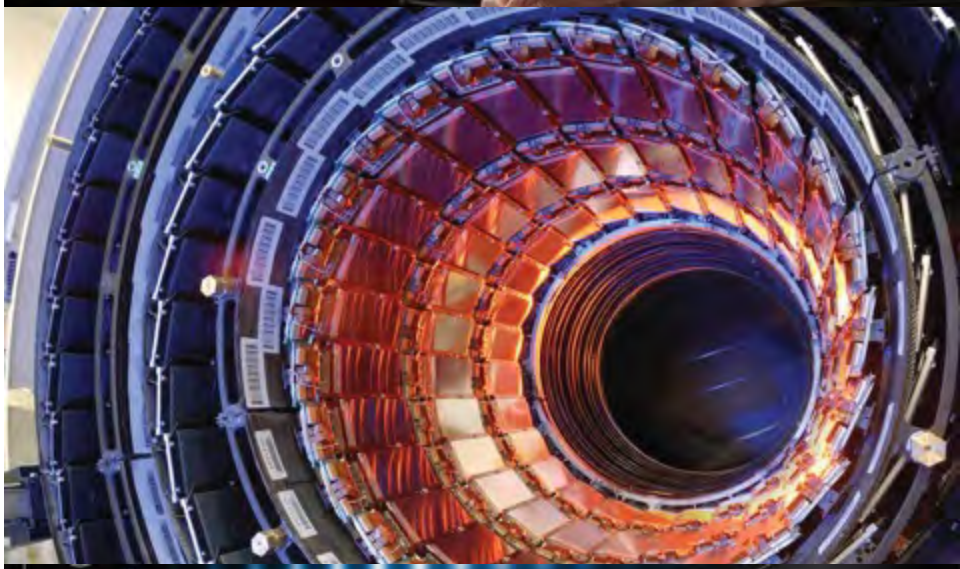
BUILT ON SAND

Steps of the Scientific Method



Validated Instruments

- Science excels at creating validated and calibrated instruments.
- Reusable & accurate
- Instrumental in reducing measurement error

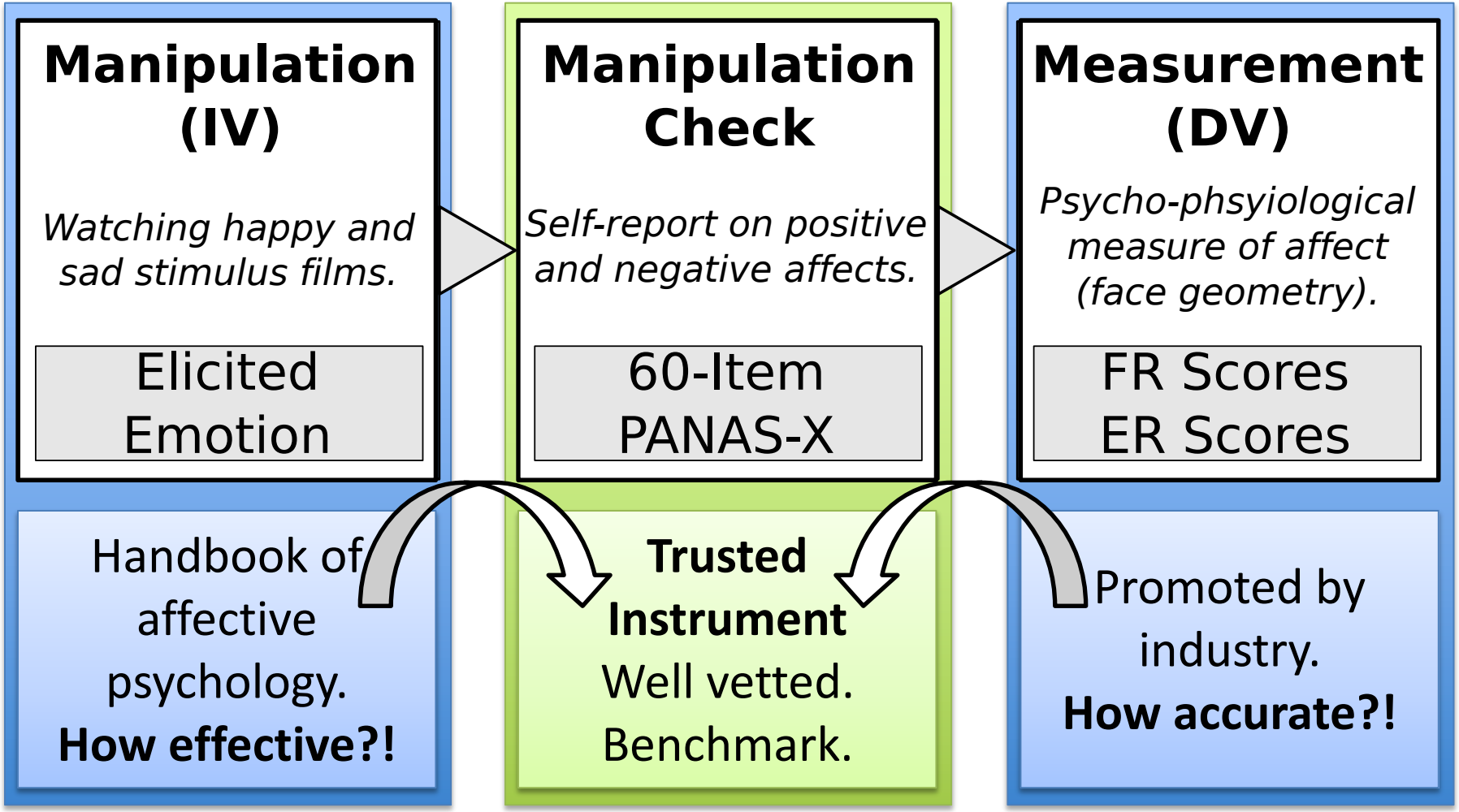


Example: Emotional Recognition

- Affect detection from facial geometry
- Compared PANAS-X Self-Report with
  MS Emotional Recognition
  NOLDUS FaceReader
- Induced Happiness & Sadness
- FR less sensitive to sadness than expected.

Example for Instrument Evaluation

Emotional Recognition

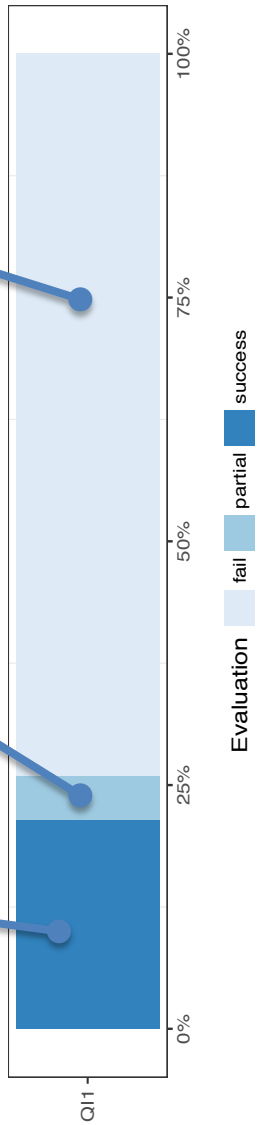


[Nwadike, Gross, Coopamootoo – Evaluating Users' Affect States… - 2016]

In security and privacy user studies, only **26%** used existing instruments.

Of a sample of **146** studies

- Three-quarter (**74%**) did not use an existing measurement or manipulation instrument.
- **5%** adapted an existing instrument, but did not follow its prescriptions precisely.
- One fifth (**21%**) replicated an existing instrument as defined.



Validated Instruments

Dos

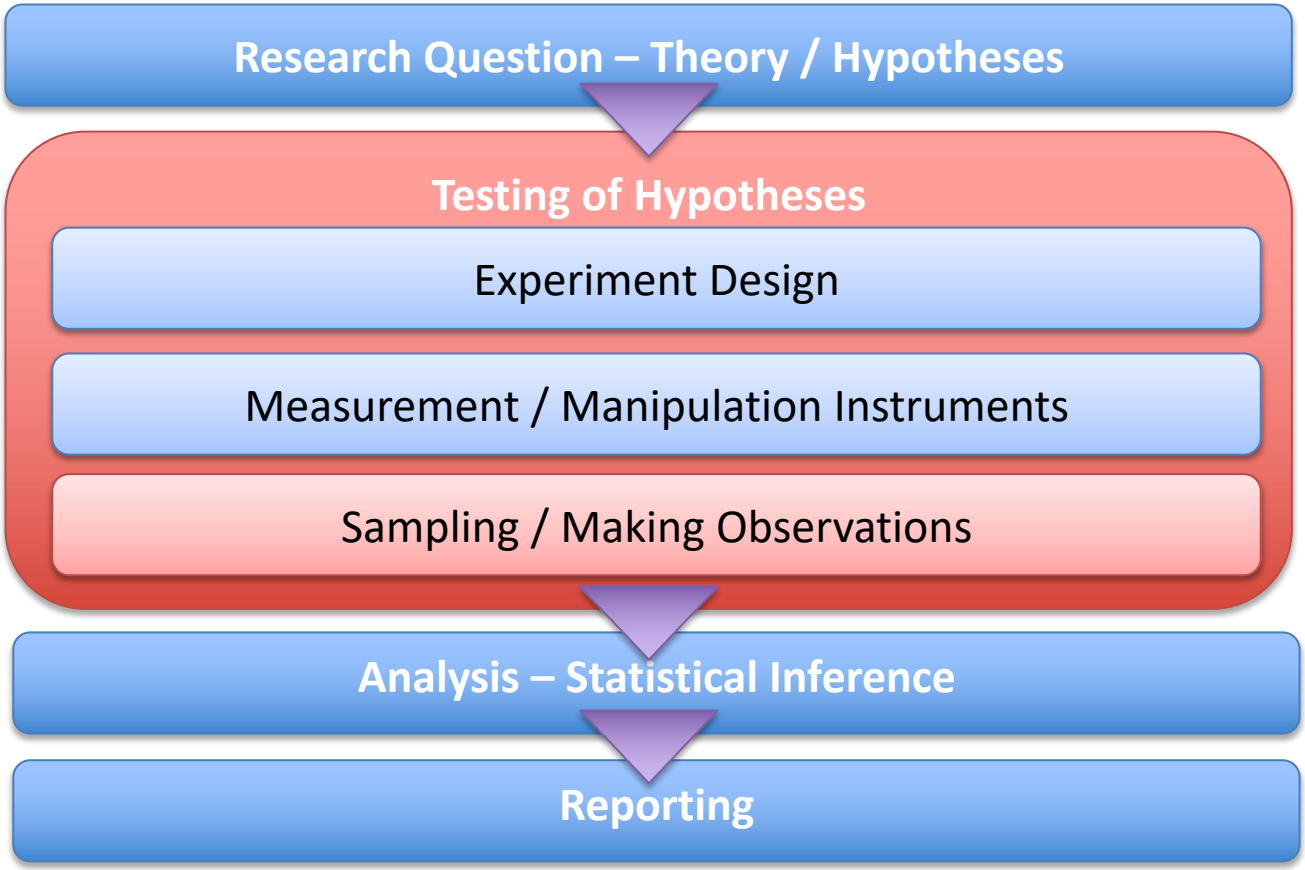
- Investigate an instrument's measured or manipulated construct and underpinning.
- Be very wary whether the instrument serves the purpose.
- Choose instruments with clear evidence of validity and reliability.
- Evaluate instruments in pre-tests yourself.

Don'ts

- Do not trust instruments just because there are popular or from an "authority."
- Do not make up your own instruments without thorough validation.
- Do not deviate from an instruments prescriptions, without re-validation.

A WARPED LENS

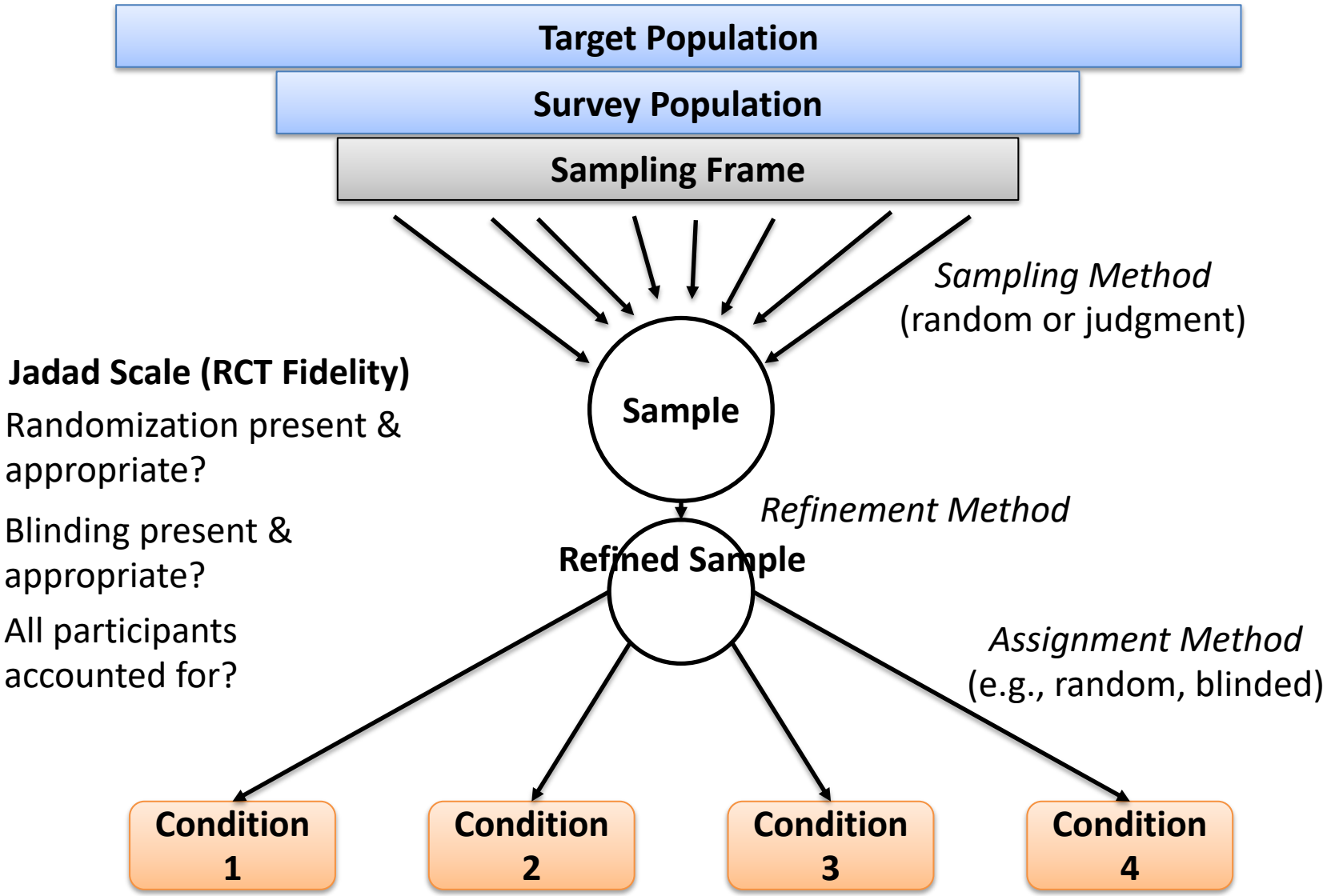
Steps of the Scientific Method



Sampling & Assignment

- Randomness can be introduced to mitigate biases effectively.
- **Random sampling** to minimize bias drawing from a population.
- **Random assignment (and blinding)** to minimize selection and experimenter biases.

Sampling and Assignment



Jadad Scale (RCT Fidelity)

Randomization present & appropriate?

Blinding present & appropriate?

All participants accounted for?

Sampling Methods

Of **146** papers in the SLR sample, **138** used quantitative methods:

- **41** (**30%**) sampled from Amazon Mechanical Turk (AMT)
- **44** (**32%**) sampled from a (largely) student population

Sampling Methods:

- **9** (**7%**) claimed simple random sampling (statistical sampling)
- **5** (**4%**) stated explicitly to use convenience sampling
- **4** (**3%**) stated explicitly to use snowball sampling
- **120** (**87%**) left sampling unspecified

Representativeness:

- **8** (**6%**) claimed to be representative of a population
- **8** (**6%**) stated *not* to be representative, but claimed generalizability to a population
- **9** (**7%**) stated explicitly not to be representative
- **113** (**82%**) did not comment on representativeness

Random Assignment and Blinding

- Few studies mentioned random assignment to conditions, fewer detailed the method used.
- Not one study employed blinding.
- Few studies fully accounted for all participants.
- **Jadad Score*** (Coded measure of RCT fidelity): At most: **2** out of **5**.

[*) Jadad Score, cf. Halpern & Douglas 2005. Appendix: Jadad scale for reporting randomized controlled trials]

Sampling & Assignment

Dos

- Explicitly define target & survey population and sampling frame.
- Prefer random sampling over judgment sampling.

- Design for double-blind.
- Use a random process for assignment to groups.
- Consider constrained random assignment.
- Report explicitly explicitly

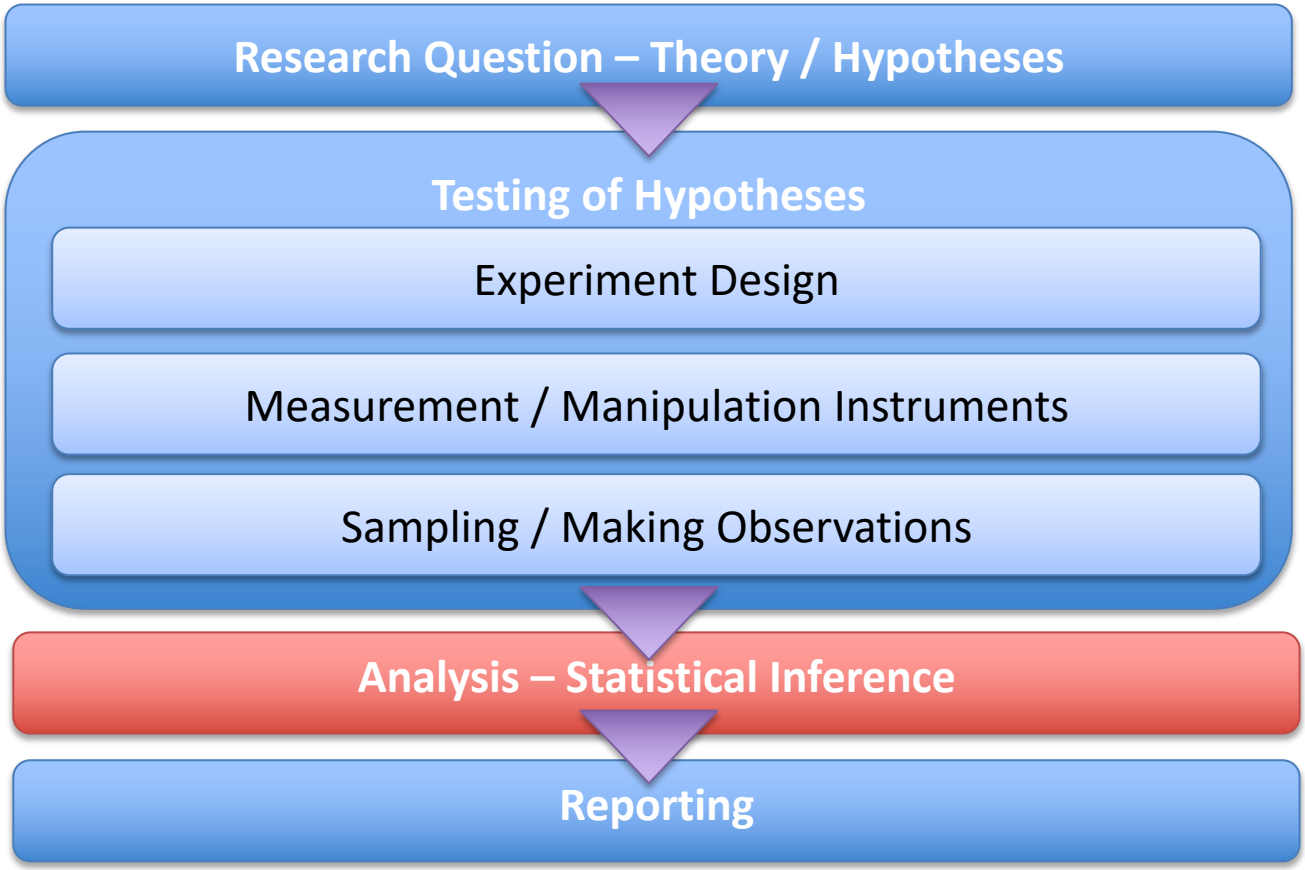
Don’ts

- Do not fall for convenience or snowball sampling.
- Do not sample from online services w/o prudent analysis of consequences.

- Never compromise on randomization for within-subject designs.
- Do not fall for processes w/o actual source of randomness (e.g., time of arrival)

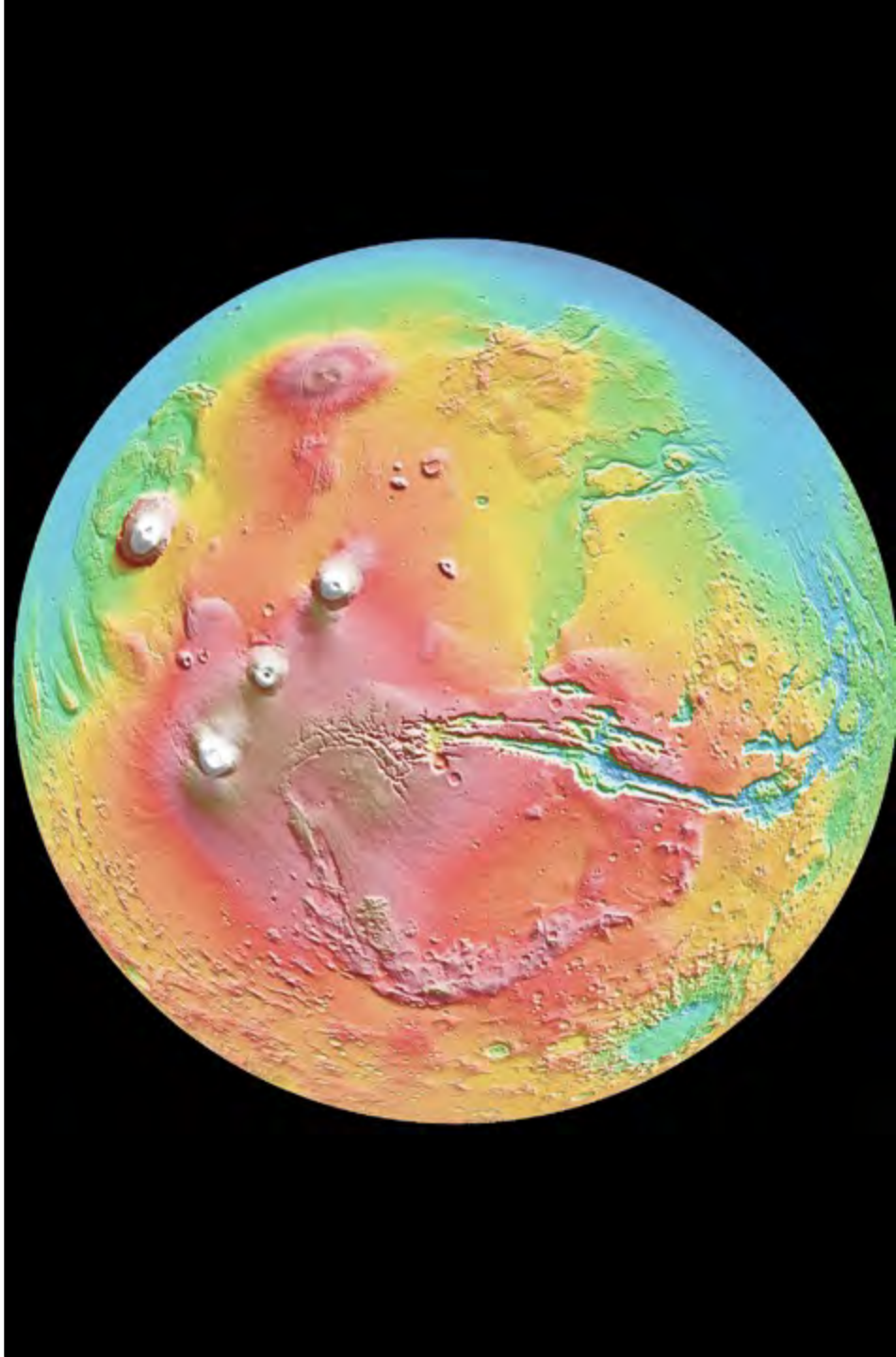
THE EARTH IS ROUND, $p < .05$

Steps of the Scientific Method



Statistical Inference

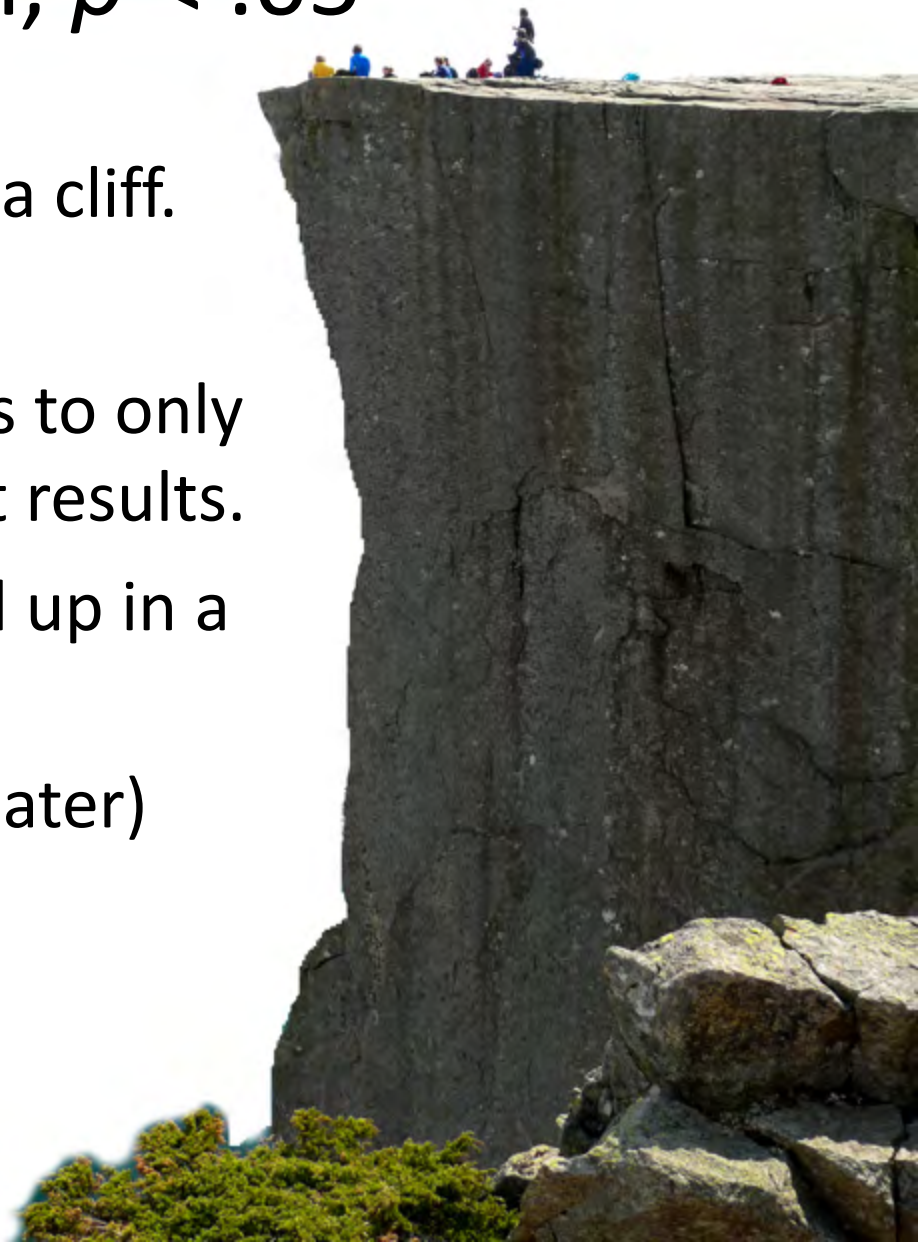
- Science shows a keen sense of estimating effects present.
- While significance testing is still common, many fields have recognized fallacies and use tools for estimation.
  - Effect Sizes
  - Confidence Intervals



The Significance Cliff, $p < .05$

- Results with $p > .05$ fall off a cliff.

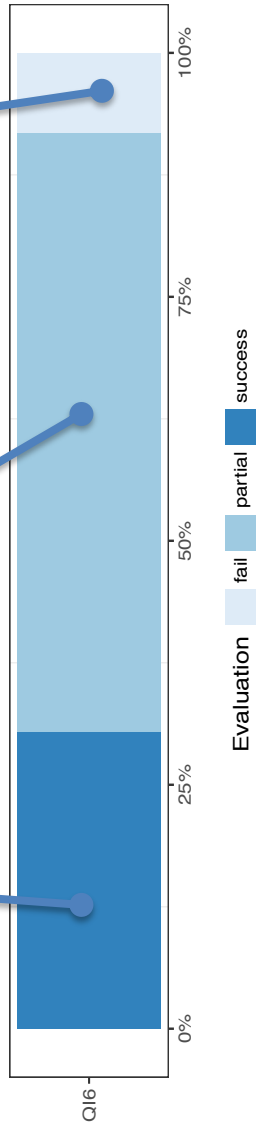
- There is a considerable bias to only publish new and significant results.
- Non-significant studies end up in a filing drawer.
- (Consider publication bias later)



Of the sample studies, only **27%** reported *p*-values with test statistics.

Of **146** sample studies

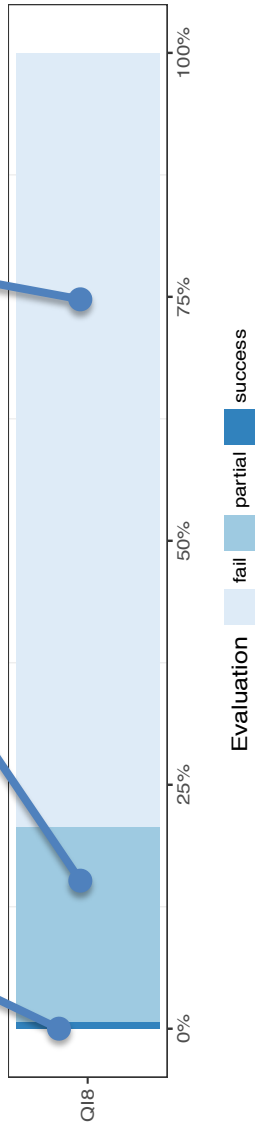
- More than two thirds (**10%**) did not offer any *p*-values/test statistics for statements made.
- About two thirds (**63%**) only partially reported test statistics and *p*-values.
- One quarter (**27%**) reported *p*-values and test statistics fully and correctly.



In security and privacy user studies, only **21%** offered effect size estimates.

Of **146** sample studies

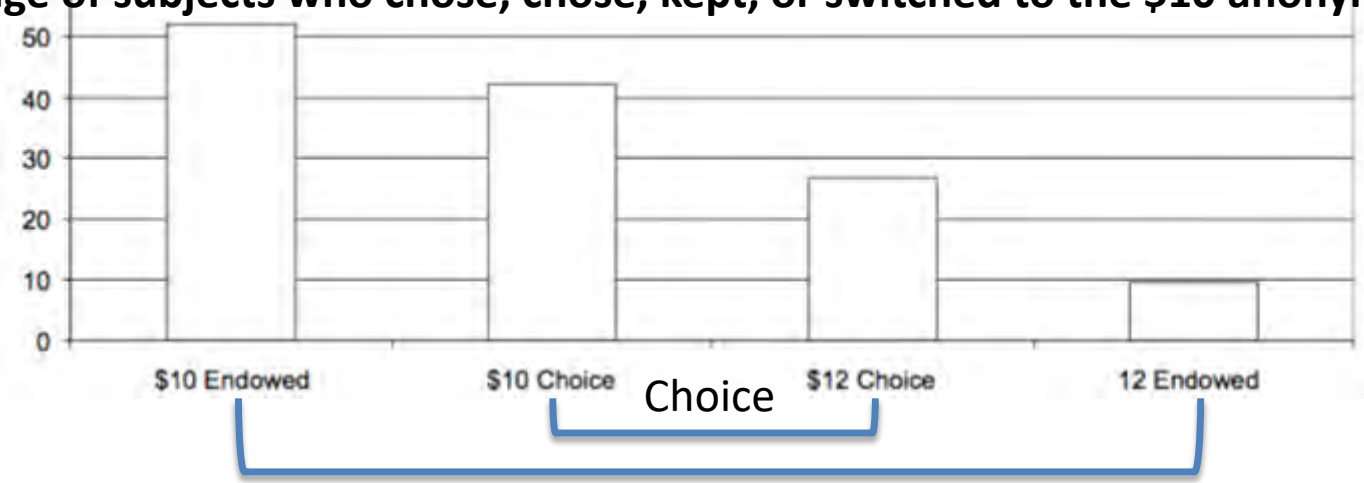
- More than two thirds (**69%**) did not offer any effect sizes or confidence intervals.
- Only one fifth (**20%**) computed effect sizes or intervals for some comparisons.
- Only **1%** computed effect sizes and their confidence intervals consistently.



“What is privacy worth?”

Through the Lens of Effect Sizes

Percentage of subjects who chose, chose, kept, or switched to the $10 anonymous card



Exposure 1:
ivate

me 1:
entifying

- **Endowed**: Participants given an anonymous $10 or identified $12 card (and need to switch from owned card)
- **Choice:** Participants can choose between anonymous $10 or identified $12 card (one of them presented first)

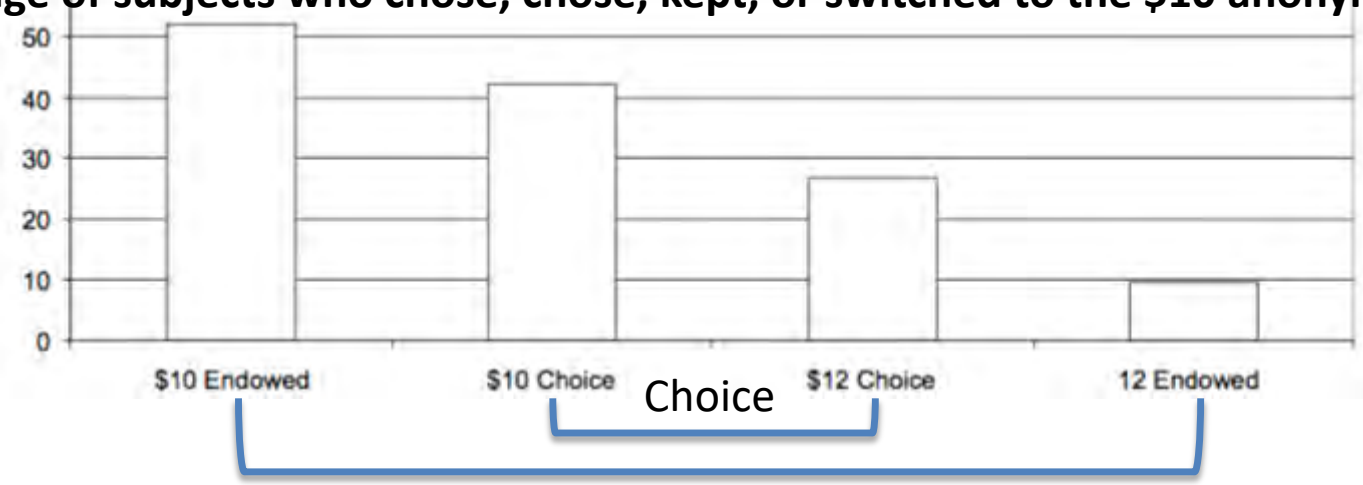
- Both effects statistically significant, $p < .05$.

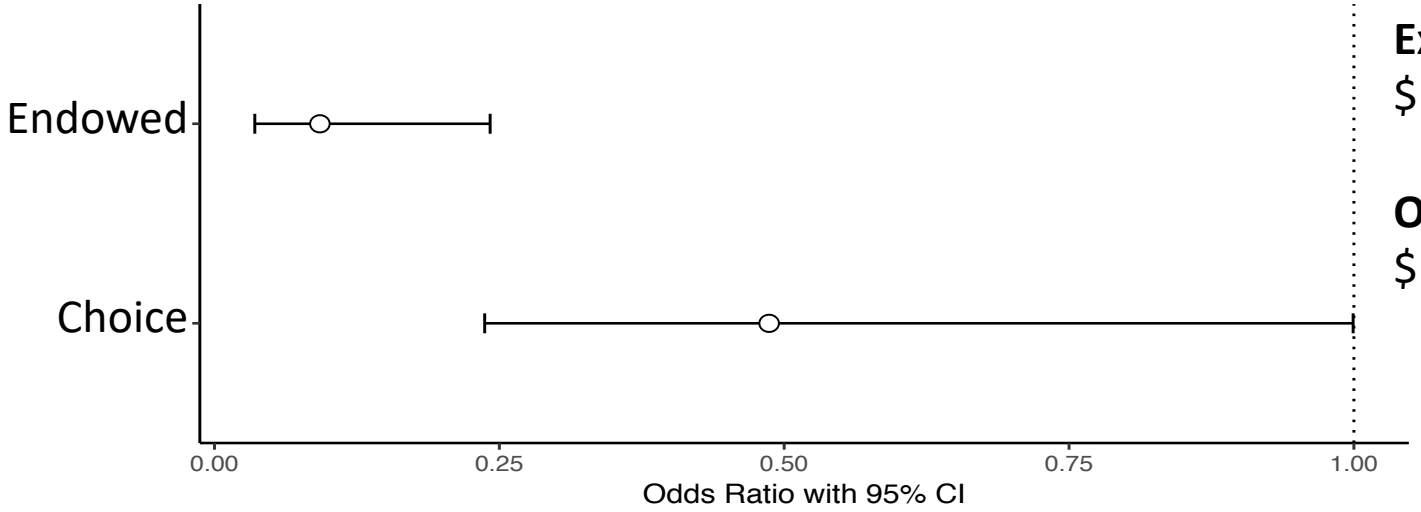
[What is privacy worth? Acquisti, John, Loewenstein, 2013]

“What is privacy worth?”

Through the Lens of Effect Sizes

Percentage of subjects who chose, chose, kept, or switched to the $10 anonymous card





Exposure 1:
$10 private

Outcome 1:
$12 identifying

[What is privacy worth? Acquisti, John, Loewenstein, 2013]

Statistical Reporting Fidelity

Reporting Triplet: (test statistic, degrees of freedom, *p*-value)

Recompute

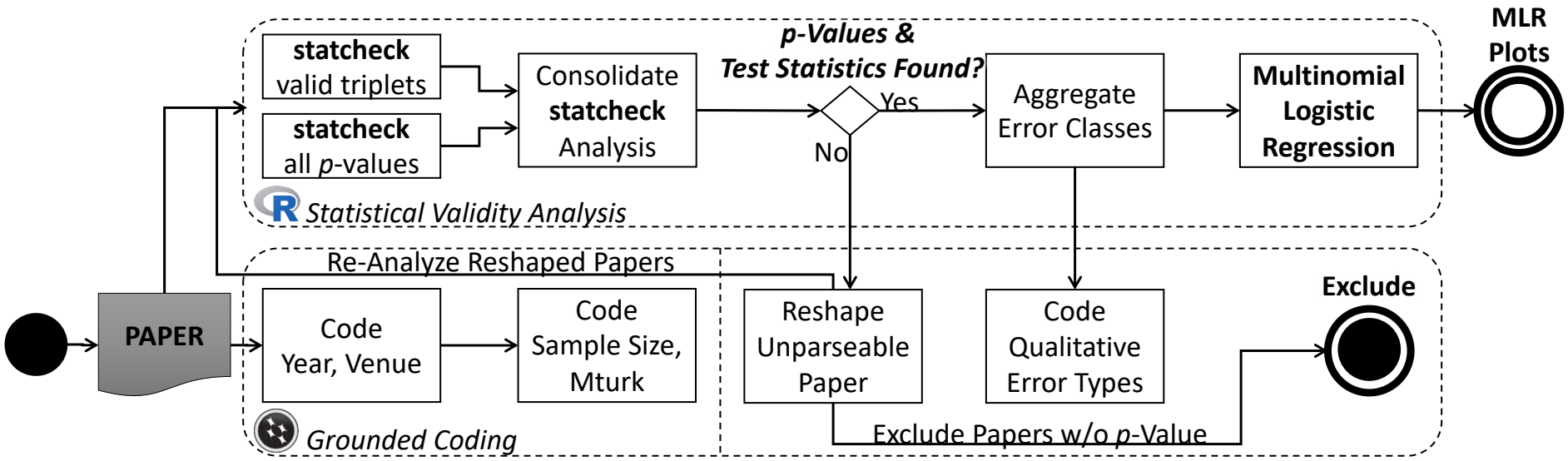
	Incomplete Triplet		Complete Triplet	
	Sig.	***p*-Value**	**ES Inferrable**	**ES Explicit**
Example	$p < .05$	$p = .019$	$t(24) = 2.52$, $p = .019$	$t(24) = 2.52, p = .019$, Hedges' $g = 0.96$, CI $[0.14, 1.76]$
***p* Quantifiable**	○	●	●	●
Cross-Checkable	○	○	●	●
ES Quantifiable	○	○	◑	●
Synthesizable	○	○	◑	●

[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Statcheck Outcome Classification

Class	Individual Test	Paper
CorrectNHST	Complete triplet reported and internally consistent: recomputed *p*-value agrees with reported *p*-value	All complete triplets in the paper are **CorrectNHST**
Inconsistency	Complete triplet reported and inconsistent: recomputed *p*-value differs from reported *p*-value	There exist complete triplets with **Inconsistency**
DecisionError	Complete Triplet reported with decision error: The recomputed *p*-value leads to a different significance decision than the reported *p*-value	There exist complete triplets with **DecisionError**
Incomplete	Incomplete triplet (*p*-value or significance comparison) reported	All reported tests are **Incomplete**.

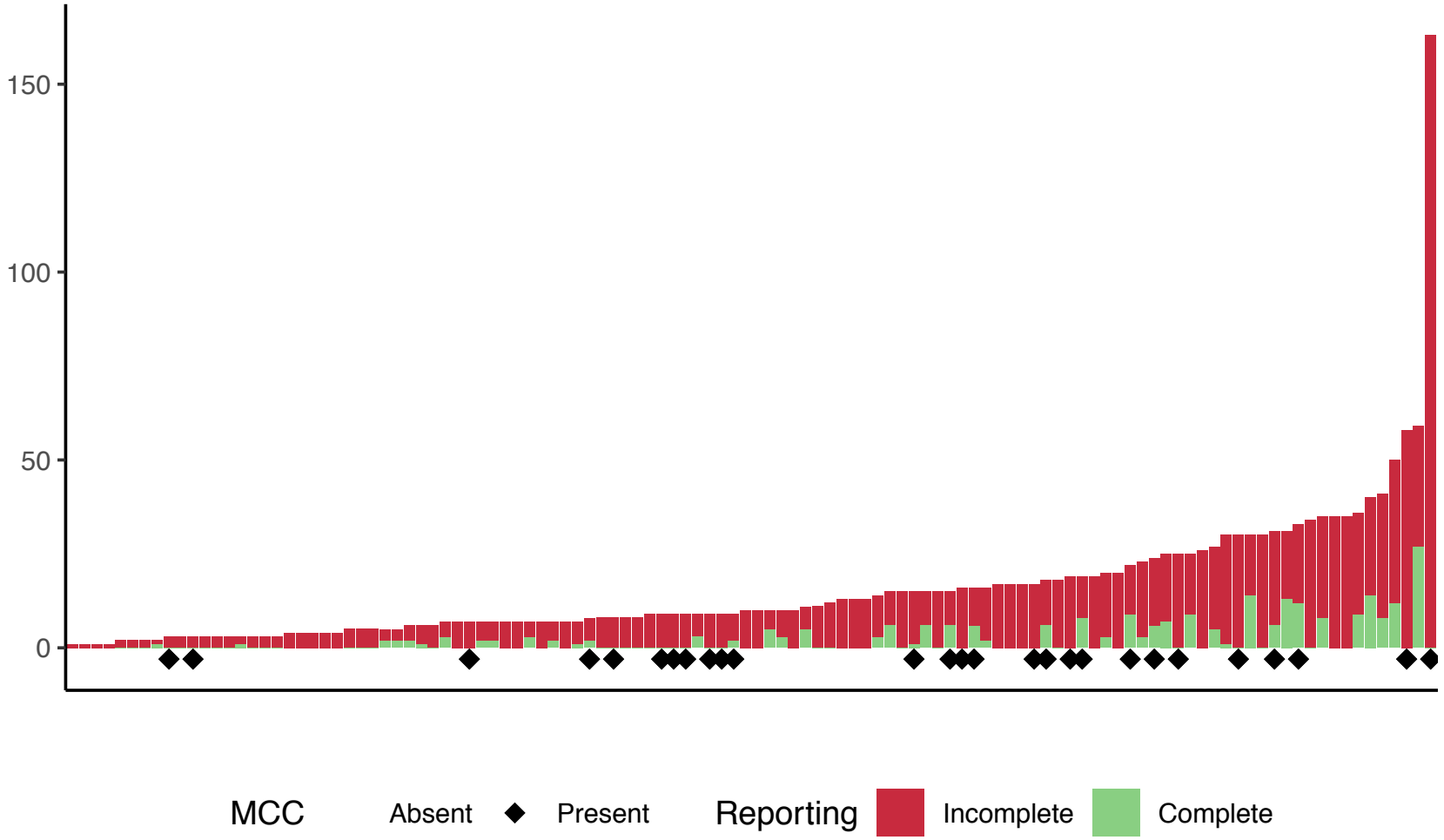
Procedure



[Preregistration available at OSF: https://osf.io/549qn/]

Distribution of Number of *p*-Values

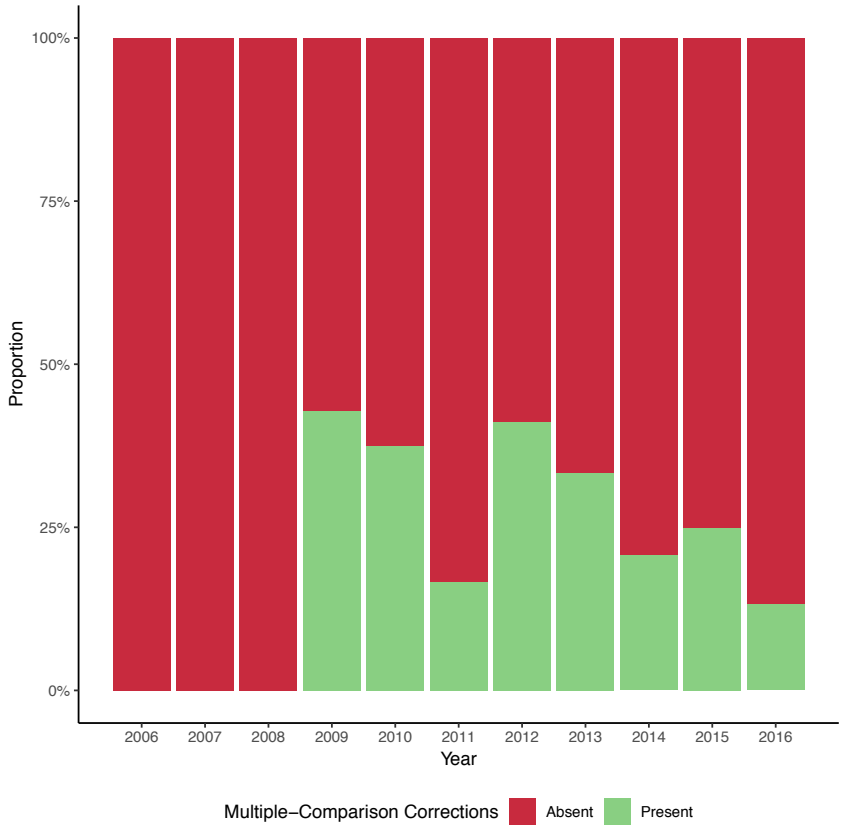
in 146 cyber security user studies

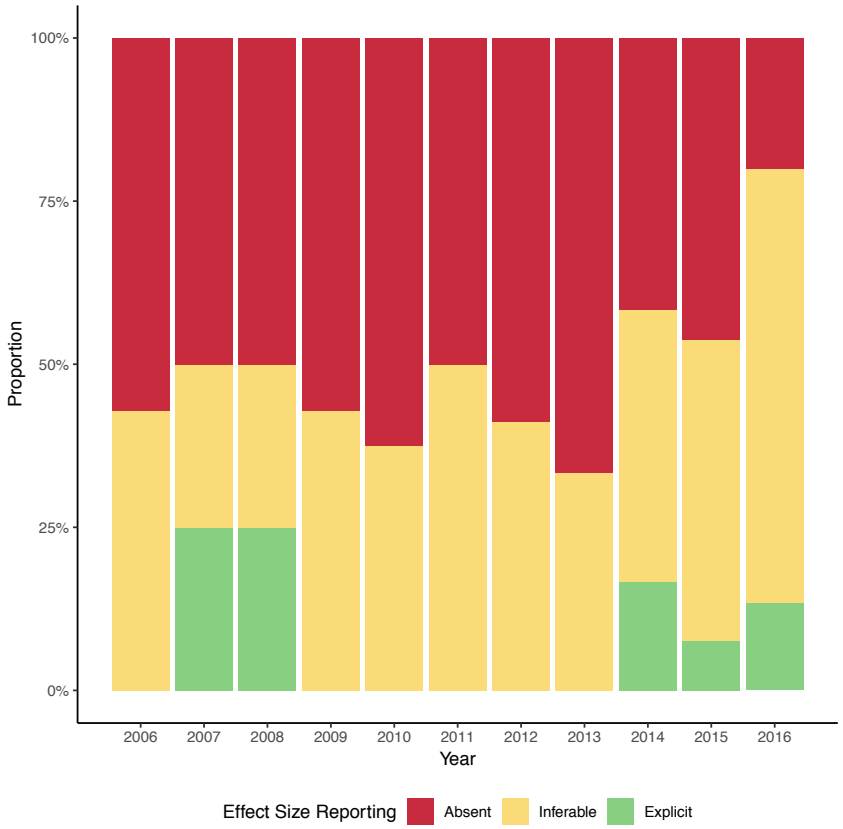


[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Per-Paper

Corrections and Effect Sizes Over Time

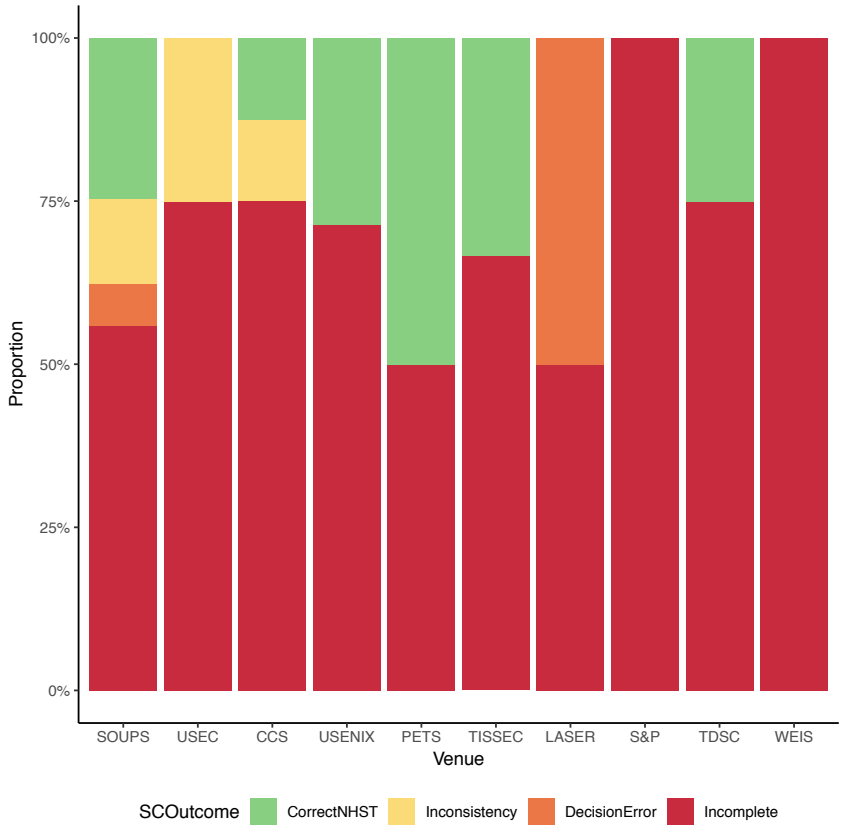


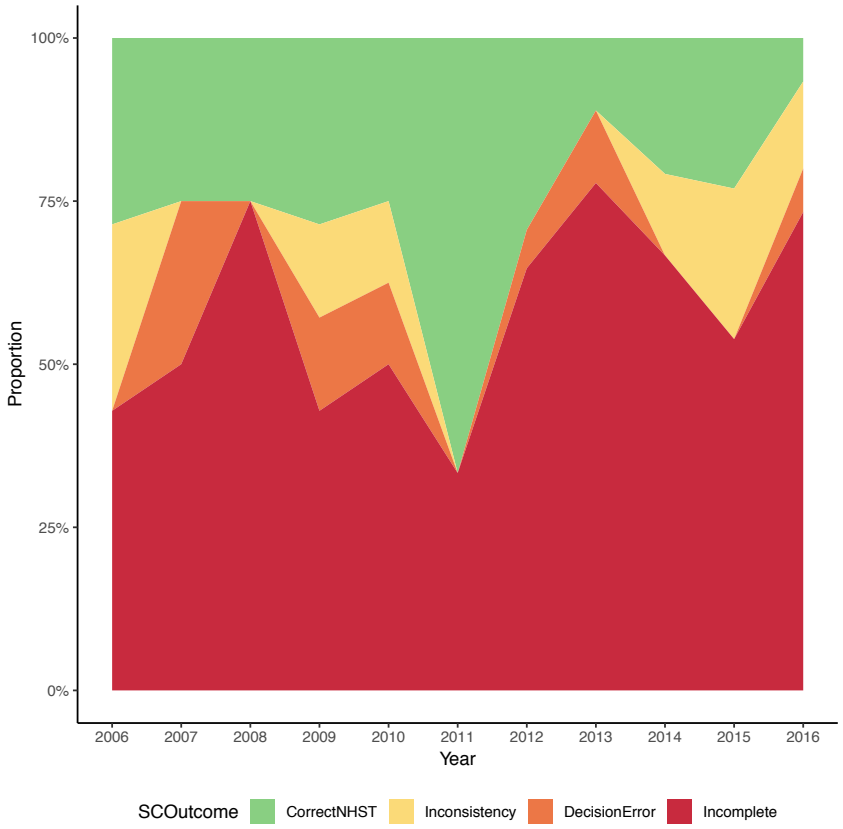


[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Per-Paper

Statcheck Outcomes by Venue & Time



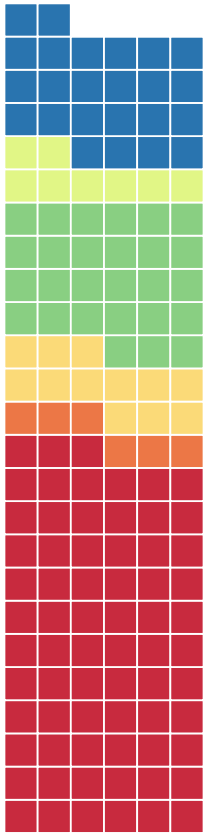


[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

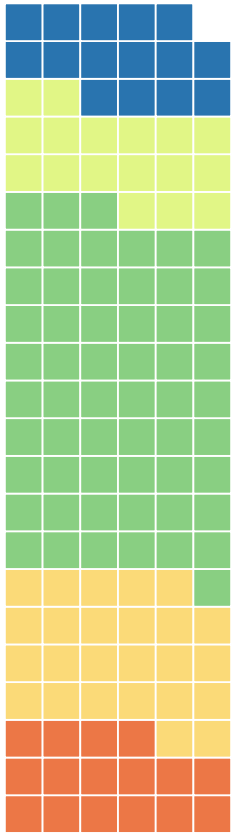
Per-Paper

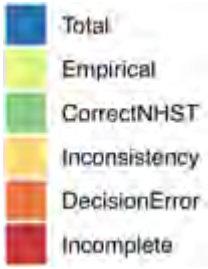
Comparison SLR vs. JMP

Cyber Security User Studies



Journal of Media Psychology

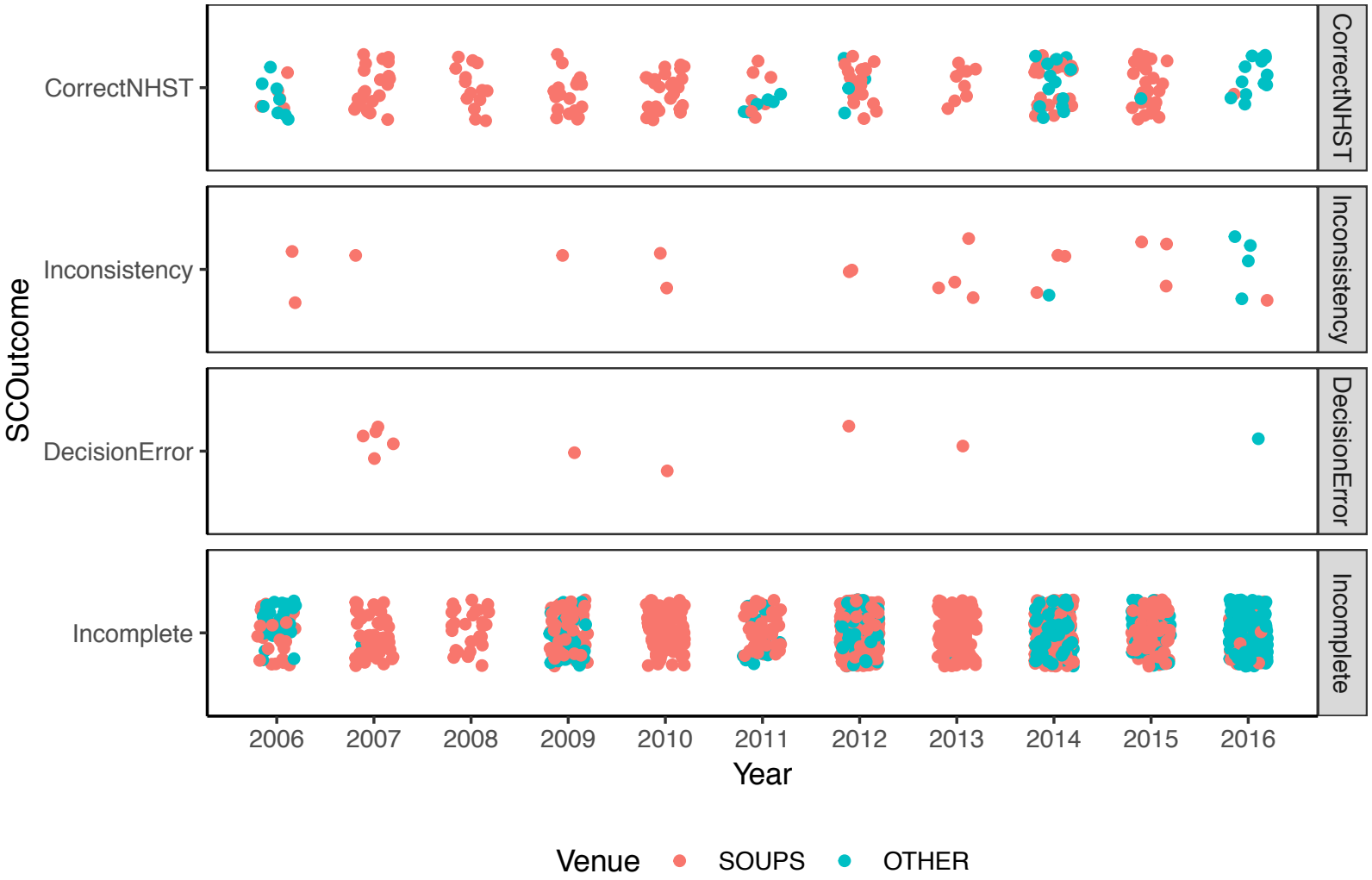




Statistically significantly different, $\chi^2(3) = 88.803$, $p < .001$
Cramer's V = 0.646, 95% CI [0.503, 0.773]

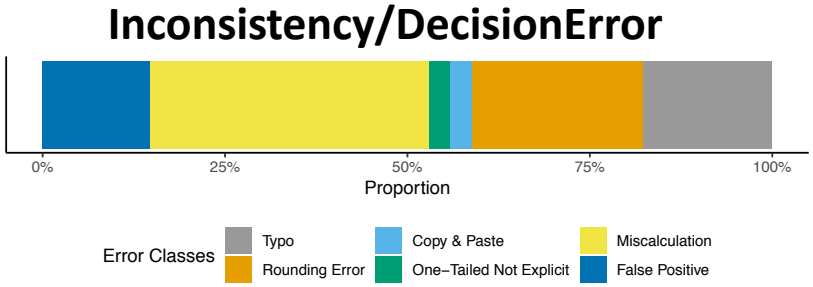
[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

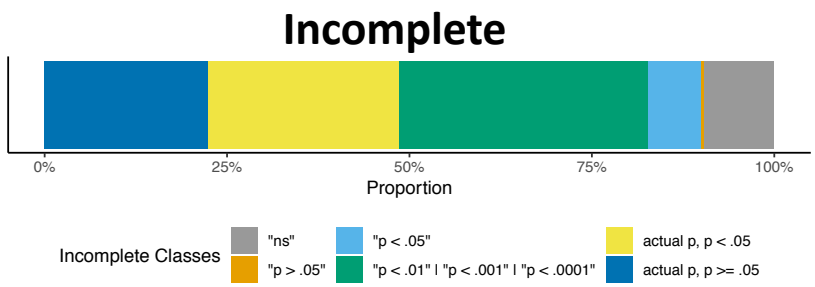
Results Over Time

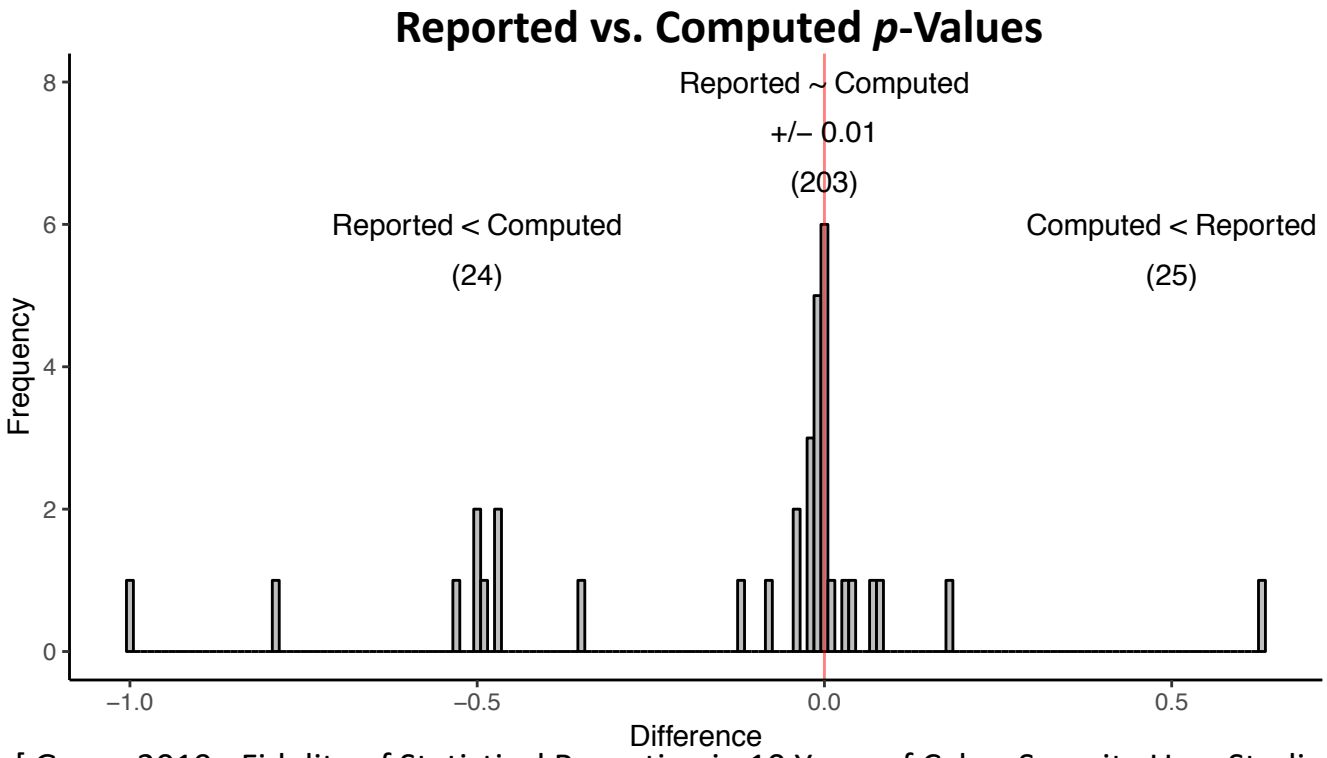


[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Qualitative Analysis







[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Discussion

- **Incomplete reporting is holding back the field**
- Nearly **two thirds** of the papers did not report a single complete test triplet.
- One quarter of the sample with complete triplets and correct statistics.
- 40% of errors were miscalculations.
- Dark figure in considerable underuse of multiple-comparison corrections.

[Gross, 2019 - Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies]

Statistical Inference

Dos

- Design with the statistical inference in mind.
- Use the right test statistic for the purpose.
- Compute meaningful effect sizes & confidence intervals.
- Use multiple-comparison corrections (MCCs).
- Report test statistics and assumptions.

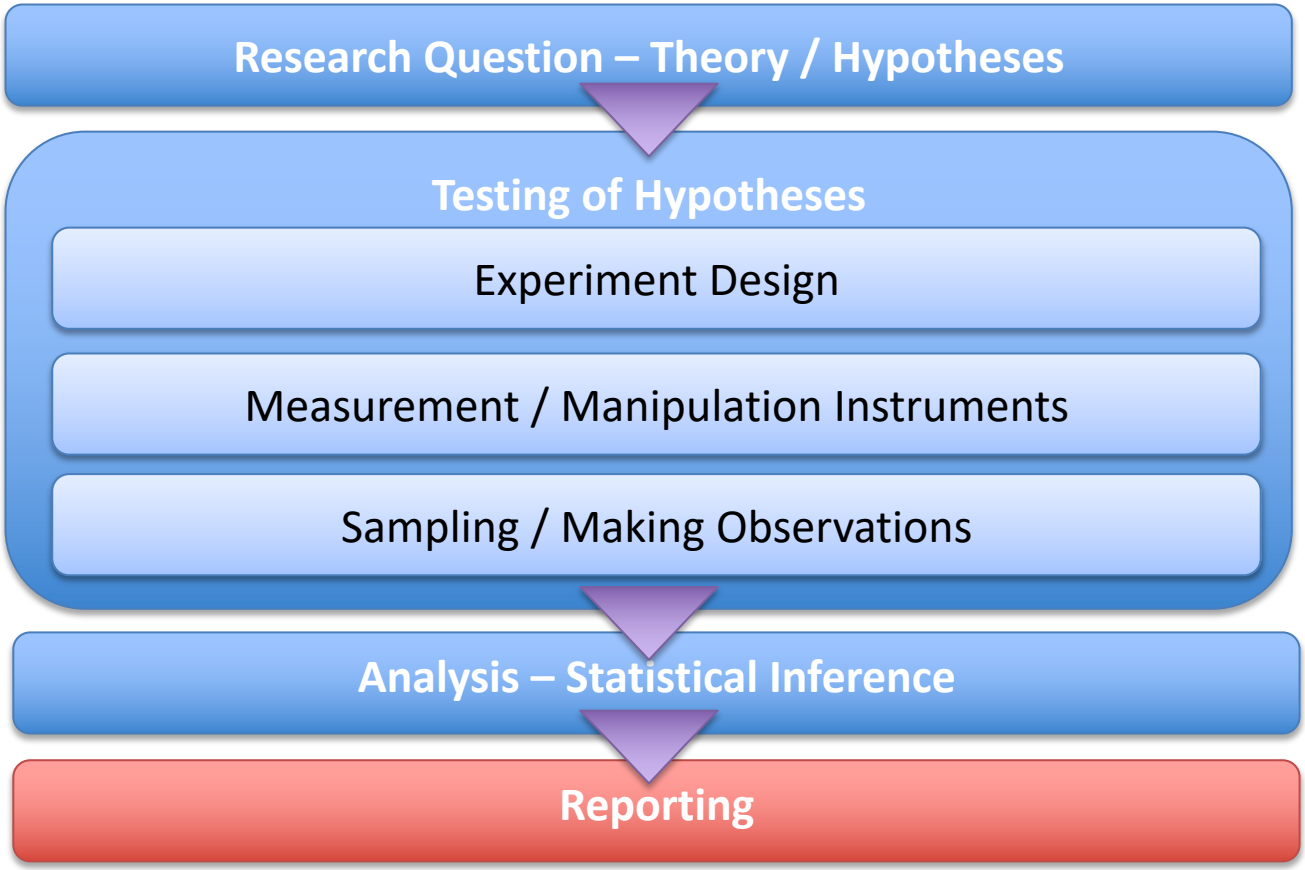
Don’ts

- Do not treat statistical as an after-thought.
- Do not allow yourself sloppiness in testing assumption and diagnostics.
- Do not violate independence assumptions.
- Do not omit inferences undertaken.
- Do not report *p*-values only.

AILMENTS OF A FIELD

CONFUSION OF TONGUES

Steps of the Scientific Method



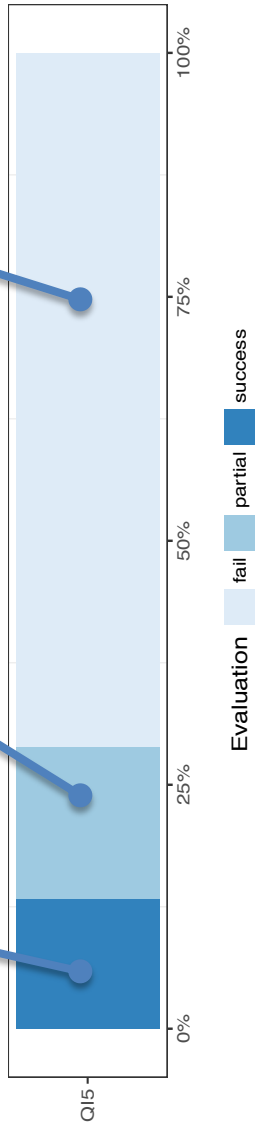
Lingua Franca of Science

- Science thrives on precise language and structured reporting.
- Overcomes ambiguity and errors by omission.
- Allows scientists and practitioners alike to communicate with high fidelity.

In user studies with human beings, only **29%** reported in APA format.

Of **146** sample studies

- **71%** did not follow the guidelines of the American Psychology Association (APA).
- **16%** followed the APA guidelines for some results.
- Less than one sixth (**13%**) followed the APA guidelines consistently.



Lingua Franca of Science

Dos

- Stick to a standardized reporting format (e.g., APA).
- Use structured abstracts and structured reports
  (Use reporting check-lists)
- Aim (incl. hypotheses)
- Method (incl. sampling, operationalization, reproducible procedure)
- Results (w/o interpretation)
- Discussion (interpretation)

Don’ts

- Do not cherry pick.
- Do not hide behind vague language.
- Do not omit information required for replication.
- Do not omit study materials.

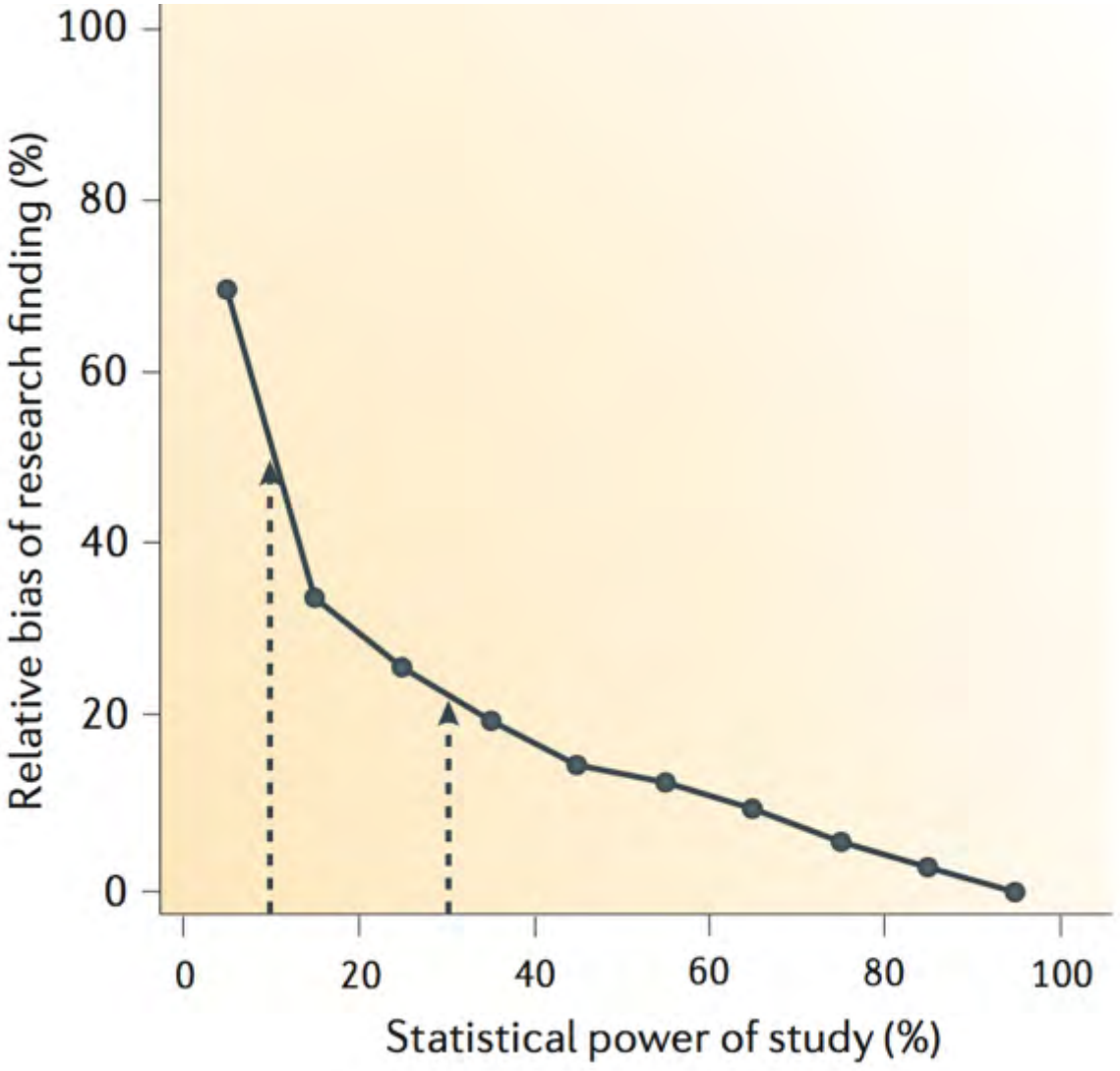
POWER FAILURE

Statistical Power

- *Statistical power* is the likelihood to detect an effect that is present in reality.
- Scientists strive to have adequate power, many aiming at **80%** or more.
- Running studies with low power, means we miss a lot of real effects.
- Low power also impacts the trust in findings.

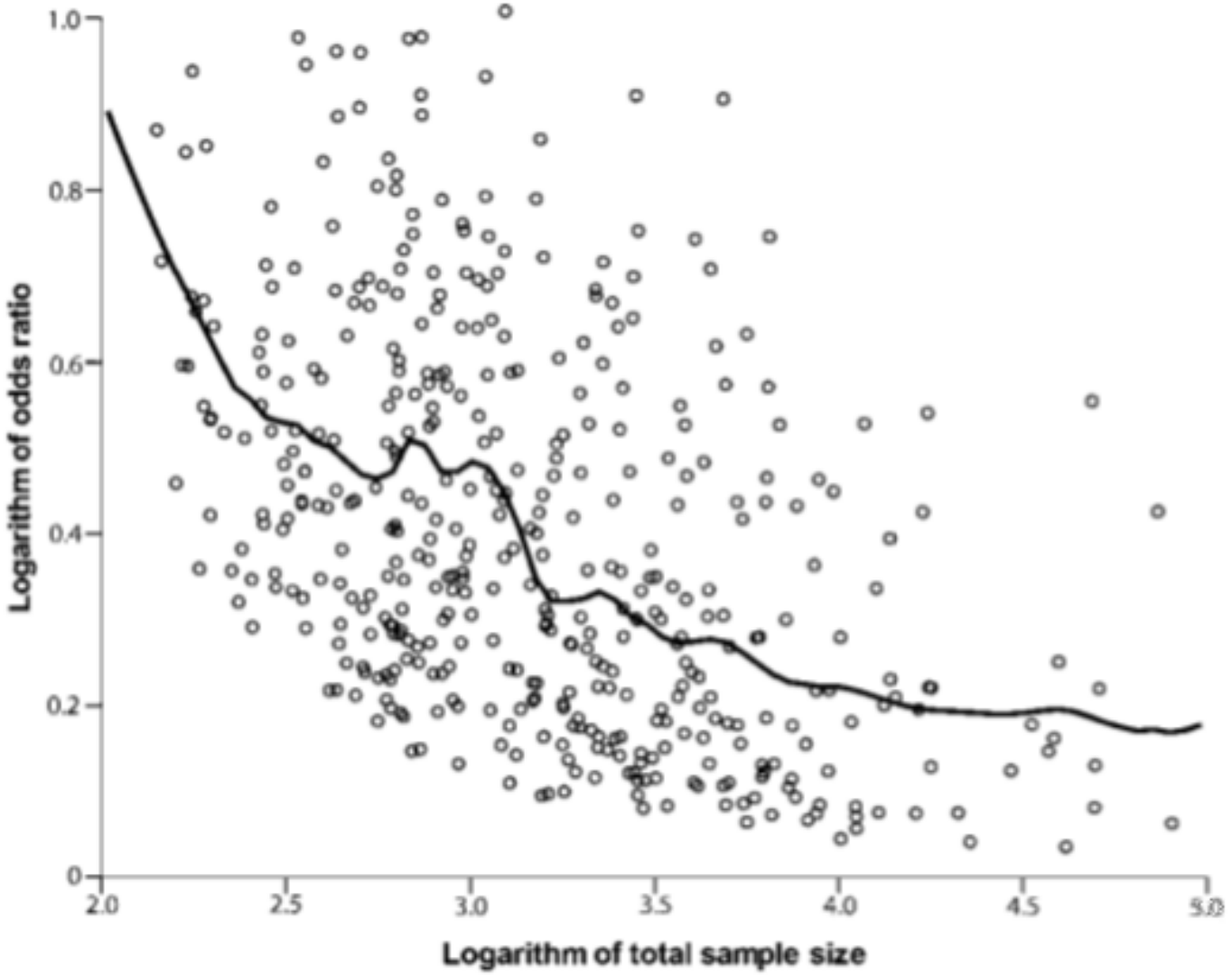
[Image from Brainbox 2013, Statistical power is...]

The Winner's Curse: Lower Power Comes With Higher Bias



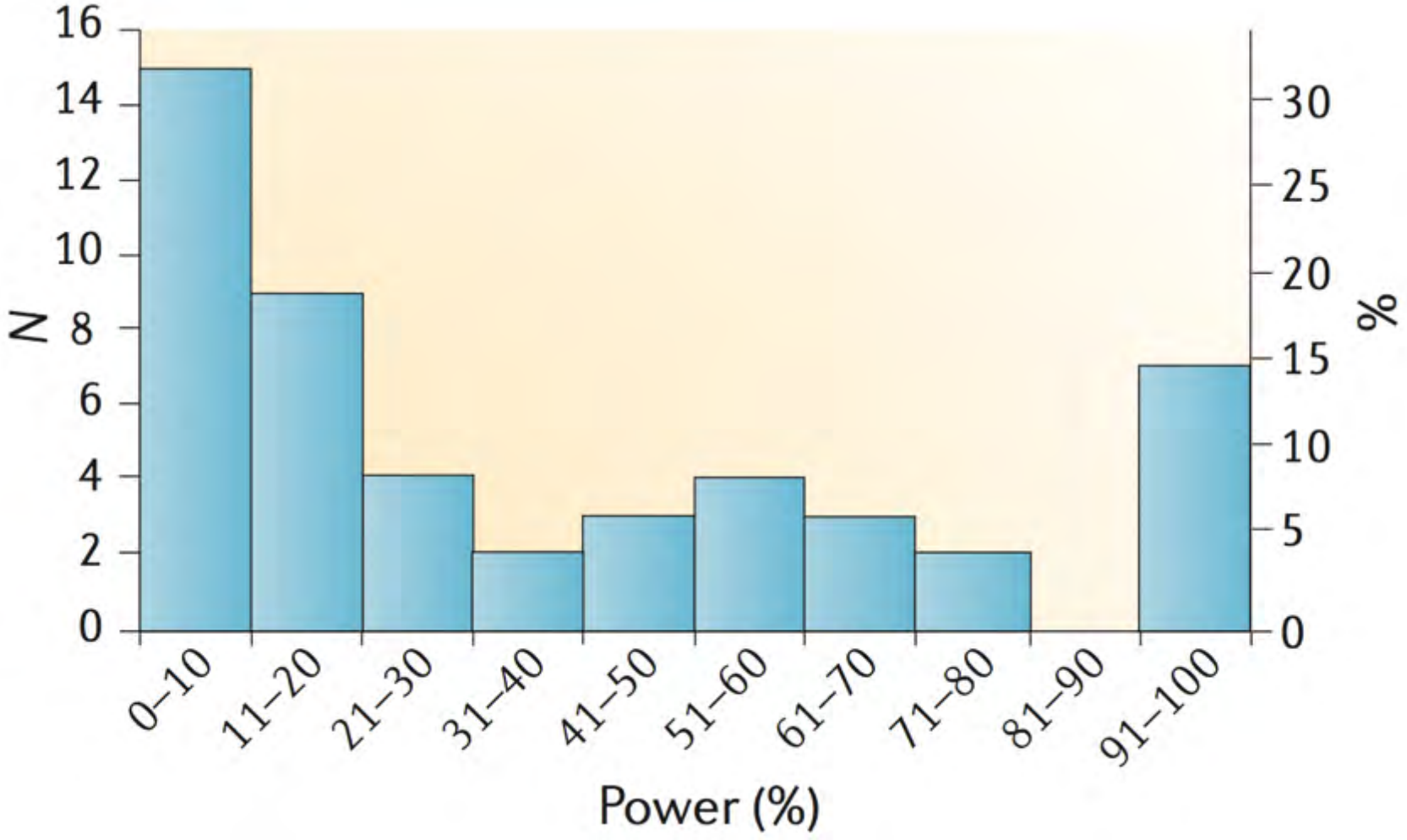
[Button et al., 2013 - Power failure, adapted from REF 103]

Low Power Inflates Effect Sizes



[Ioannidis 2008, Why most discovered true associations are inflated]

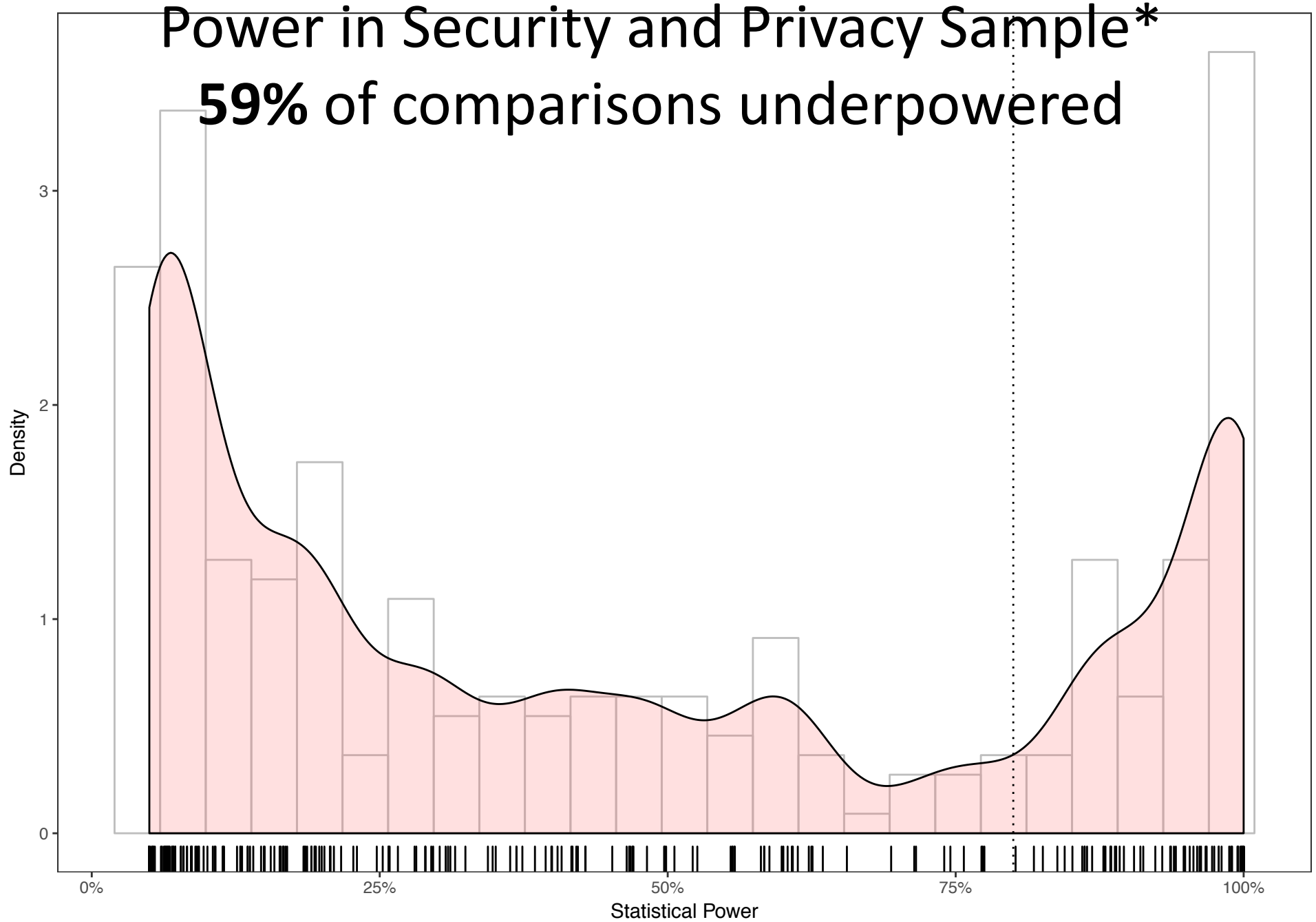
Median Power in Sample of Neuroscience Studies



[Button et al., 2013 - Power failure]

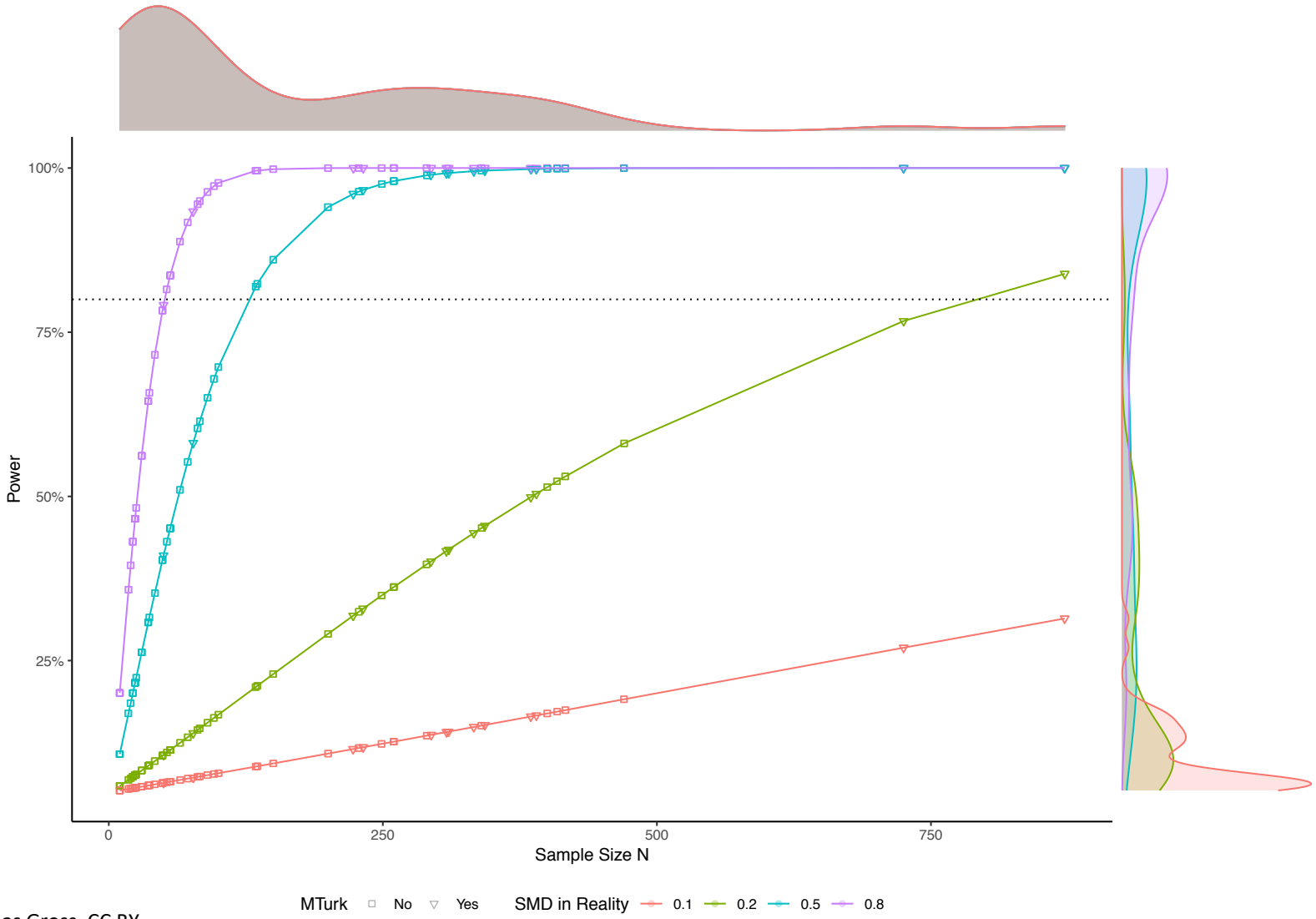
Power in Security and Privacy Sample*

59% of comparisons underpowered



[*) From the quantitative sample of our SLR, 19 studies, 277 comparisons]

Power Simulation
Wrt. Parametrized Effect Sizes



"Why most published research findings are false."

John P. A. Ioannidis

- The power failure plays a major part in "positive" reports being false positives.
- The lower the power, the less likely reported findings are true in reality.
- Poorly performed studies accumulating a bias are especially prone to yield false positives.

Positive Predictive Value

Positive Predictive Value (PPV): The likelihood of the alternative hypothesis $\mathbf{H_1}$ being true in reality, given the experiment's observation *D*.

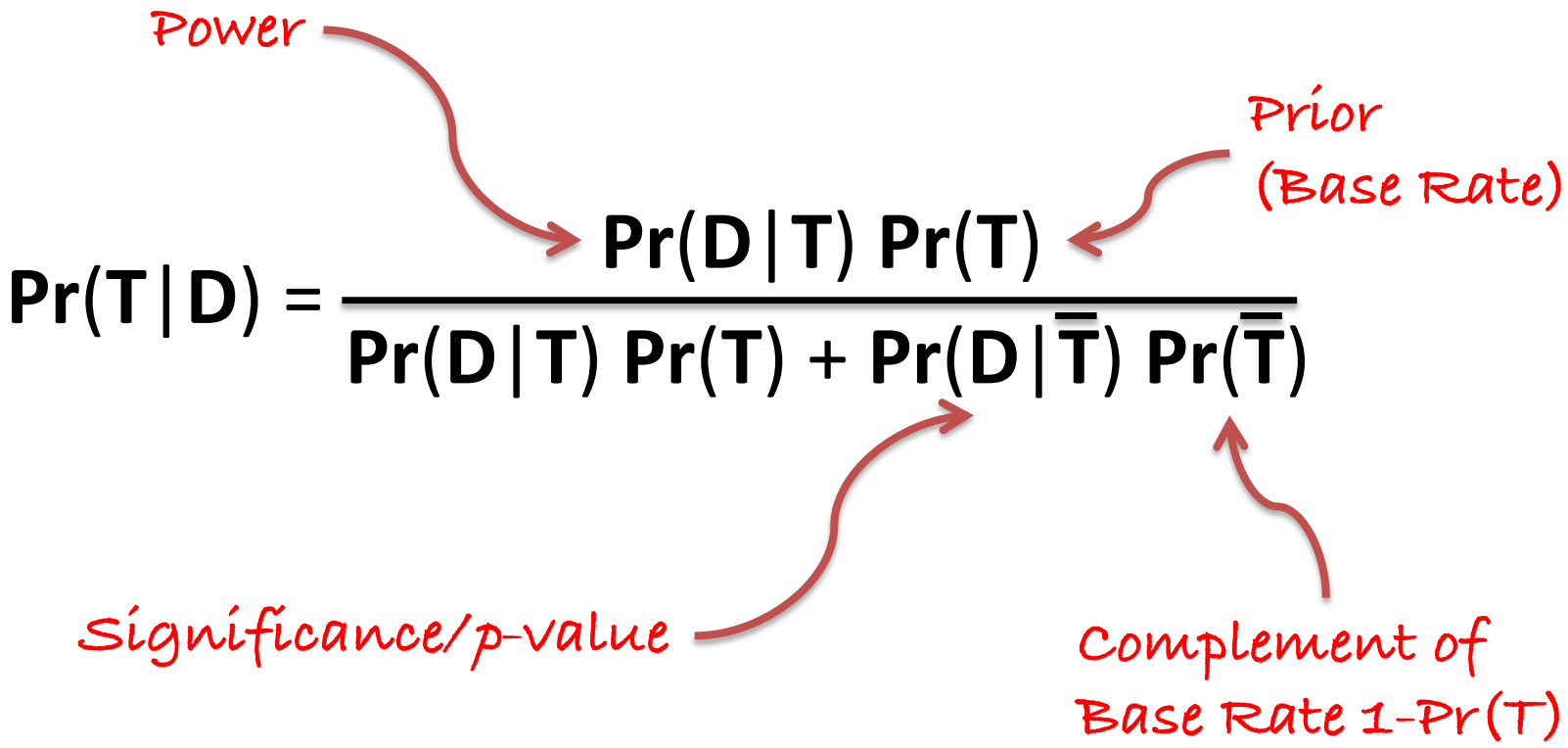
Derived with Bayes Theorem. Impacted by:

- Significance level α
- Power 1-β
- Prior likelihood of alternative hypothesis R=$\mathbf{Pr}(\mathbf{H_1})$
- Experimental bias *u* (Estimate cf. Ioannidis)

 [Ioannidis – Why Most Published Research Findings Are False]

Deriving the Positive Predictive Value

Via Bayes Theorem



T: True Relationship, $\mathbf{H_1}$

D: Observation/Data D

Typical Form of Empirical Cyber Security User Studies?

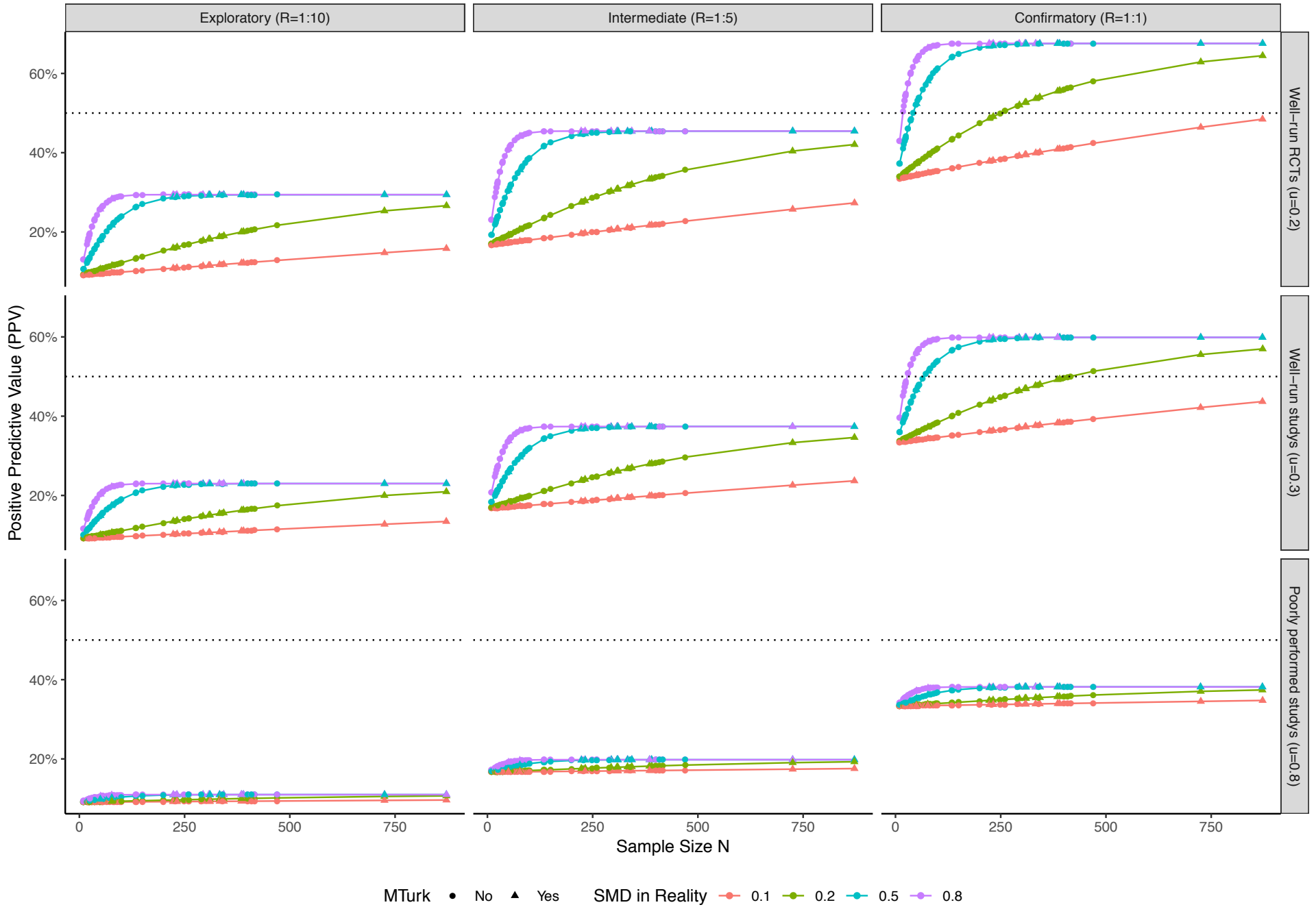
Typical Prior Probability

Typical Bias	Exploratory (R=1:10)	Intermediate (R=1:5)	Confirmatory (R=1:1)
Well-run RCTs	Well-run exploratory RCTs	Well-run intermediate RCTs	Well-run confirmatory RCTs
Well-run studies	Well-run exploratory studies	Well-run intermediate studies	Well-run confirmatory studies
Poorly performed studies	Poorly performed exploratory studies	Poorly performed intermediate studies	Poorly performed confirmatory studies

[Interactive exercise of the talk, replaced with a static slide]

Simulation of PPVs

- The subsequent slide contains the results on simulation of PPV, depending an assumed mean effect size in the field and on parametrized mean prior probability of a field and the typical bias.
- **Recall:** PPV is the probability of a positive report being true in reality.
- **Bottom line:** Only high-powered, well-run confirmatory studies reach a posterior probability greater than 50%, that is, better than a coin flip.



Statistical Power

Dos

- Estimate prior probability; relationships likely true.
- Estimate likely effect sizes.
- Have sound *a priori* power analysis (e.g., G*Power*).
- Pre-empt multiple-comparison corrections
- Consider Accuracy in Parameter Estimation (AIPE**)

Don’ts

- Do not start a study blue-eyed, hoping for the best.
- Do not ignore recommended minimal group sizes.
- Do not compromise on statistical power.

[*) G*Power: Faul & Erdfelder 2007. G* Power 3: A flexible statistical power analysis …]
[**) AIPE: Maxwell, Kelley, Rausch 2008. Sample size planning for statistical power and accuracy in parameter estimation]

SCIENTISTS BEHAVING BADLY

Scientific Integrity

- A moral compass of scientists is essential in guaranteeing a field's integrity.
- Scientists are seeking to root out a variety of possible questionable research practices (QRPs).
- Though often committed without bad intentions, QRPs can undermine the reliability of results.

High-Impact Questionable Research Practices

Outcome Switching*

- Claiming to have predicted an unexpected finding.
- Substituting an unplanned yet significant hypothesis for a planned aim/hypothesis.

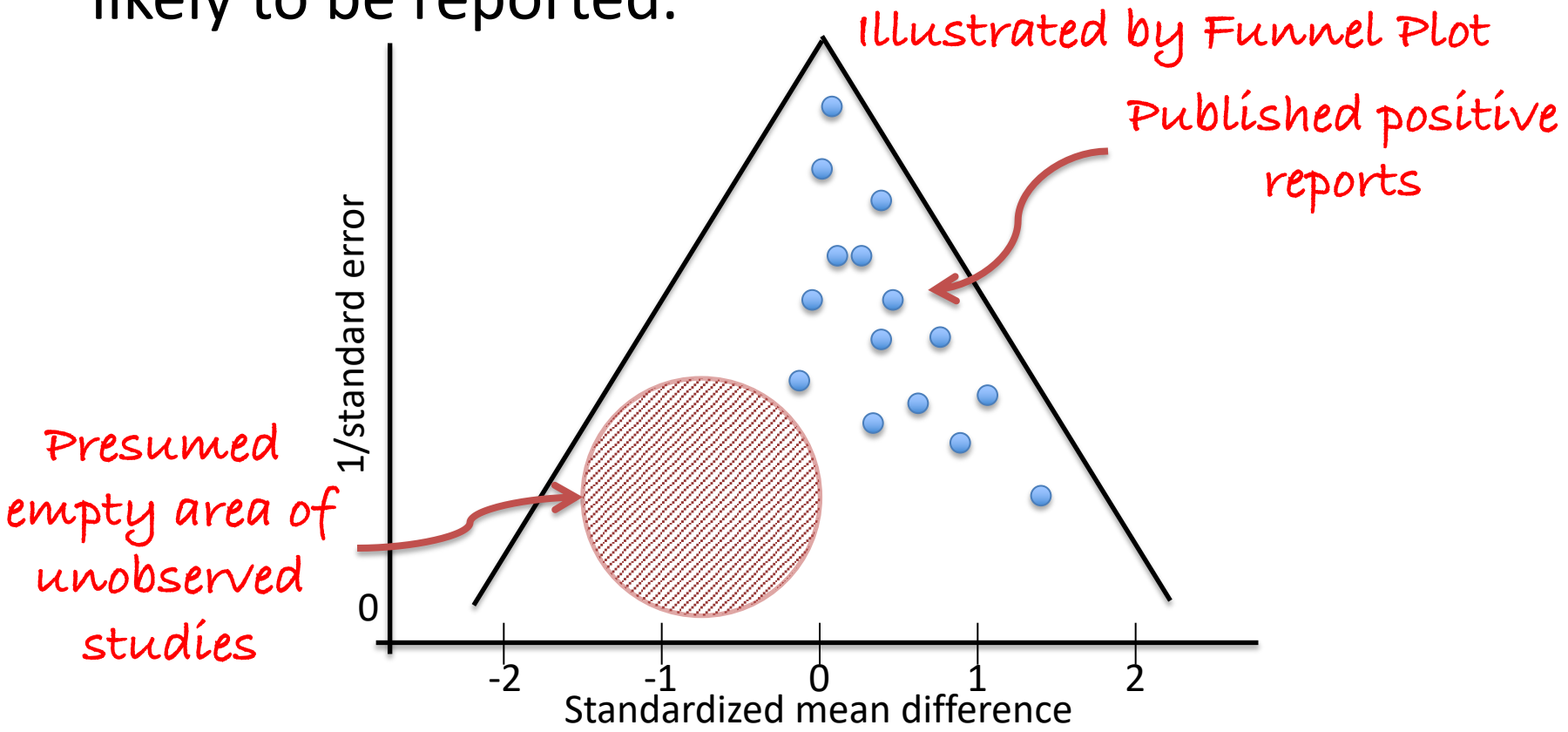
Data Dredging (*p*-Hacking)**

- Trawling the data for significant hypotheses.
- Testing many models/variable combinations till a significant pattern emerges.
- Reporting only the significant results.

[*) Altman, Moher, Schulz 2017. Harms of outcome switching in reports of randomised trials: CONSORT perspective]
[**) *p*-Hacking: Lead et al. 2015. The Extent and Consequences of P-Hacking in Science]

A Word on the Publication Bias

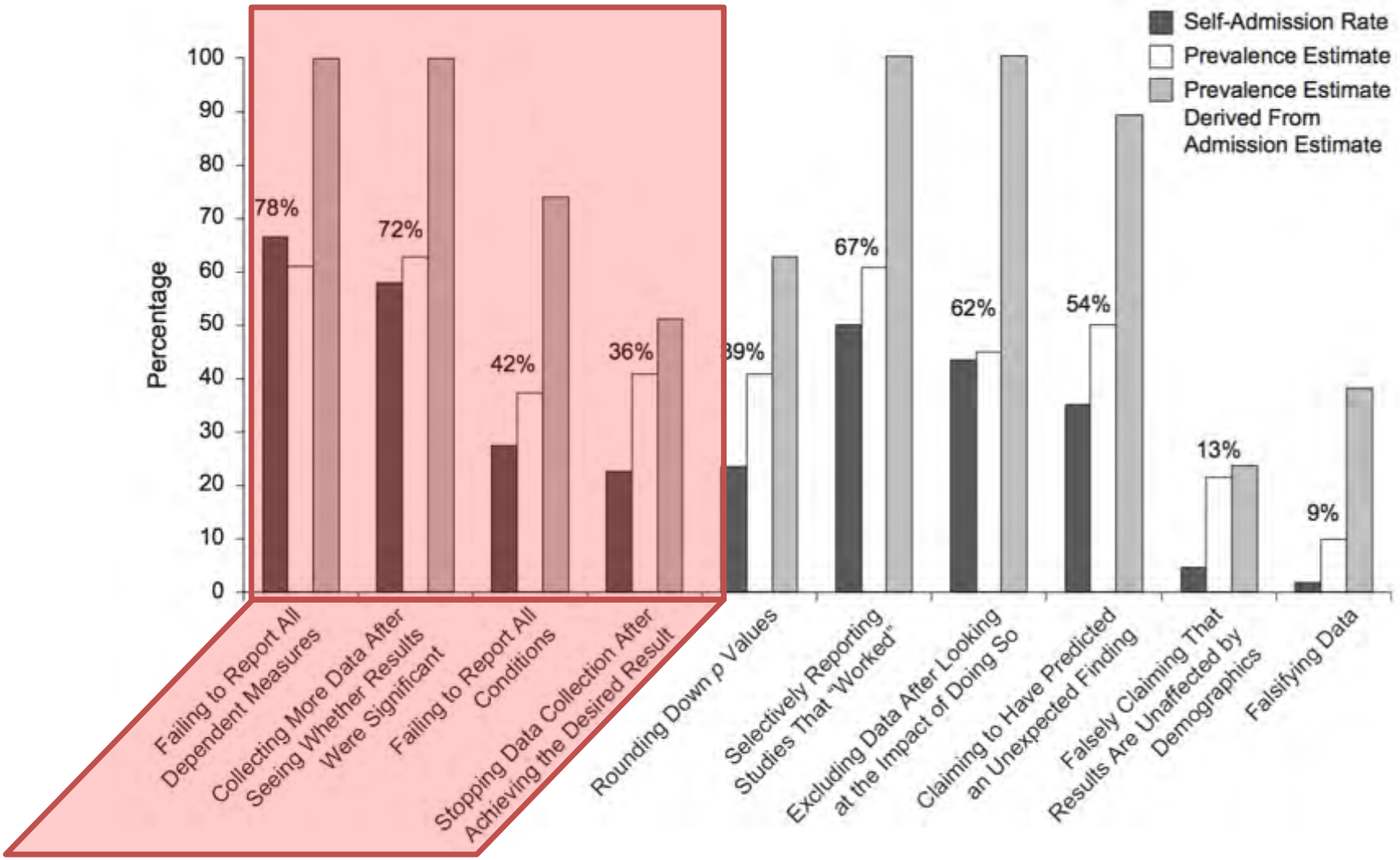
Publication Bias: Outcome of a study influences the likelihood of publication; positive findings are more likely to be reported.



[Adapted from Duval, Tweedie – Trim and Fill – A Simple Funnel-Plot-Based Method… – 2000]

Questionable Research Practices Prevalence in Psychology

Let's have a closer look!



[John, Loewenstein, Prelec: Measuring the Prevalence of Questionable Research Practices, 2012]

Flexibility Breeds False Positives

Looking at data through multiple lenses, considering multiple facets, amplifies the risk of false positives.

- Comparisons made.
- Conditions included or not.
- Covariates included or not.
- Testing data while collecting.

Flexibility in False-Positive Psychology

Flexibility in data collection, analysis and reporting inflates actual false positive rates.

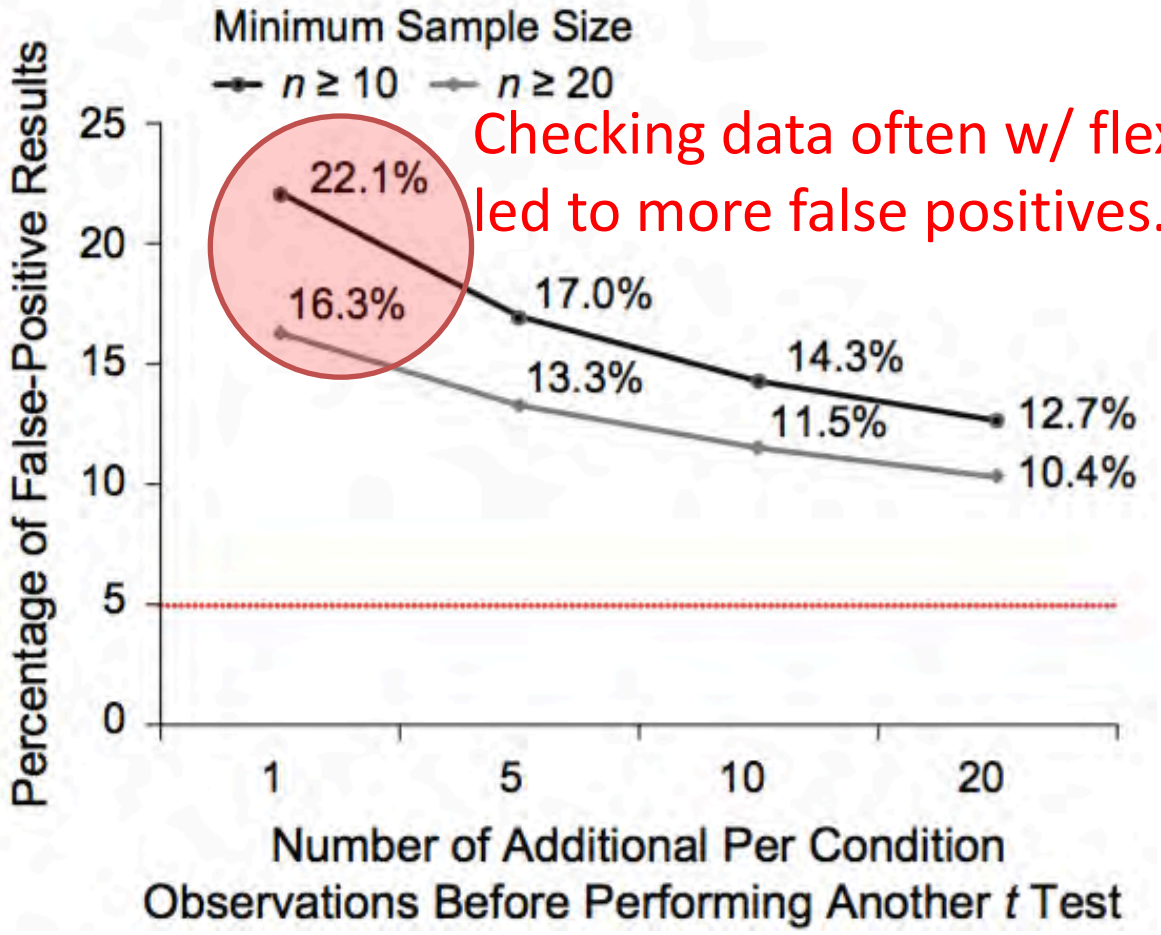
Researcher degrees of freedom	Significance level $p < .05$
Situation A: two dependent variables ($r = .50$)	9.5%
Situation B: addition of 10 more observations per cell	7.7%
Situation C: controlling for gender or interaction of gender with treatment	11.7%
Situation D: dropping (or not dropping) one of three conditions	12.6%
Combine Situations A and B	14.4%
Combine Situations A, B, and C	30.9%
Combine Situations A, B, C, and D	60.7%

Exploiting flexibility led to more false positive "significant" findings.

[Simmons, Nelson, Simonson: False-Positive Psychology, 2011]

Flexibility in False-Positive Psychology

Testing during data collection inflates false positive rates.



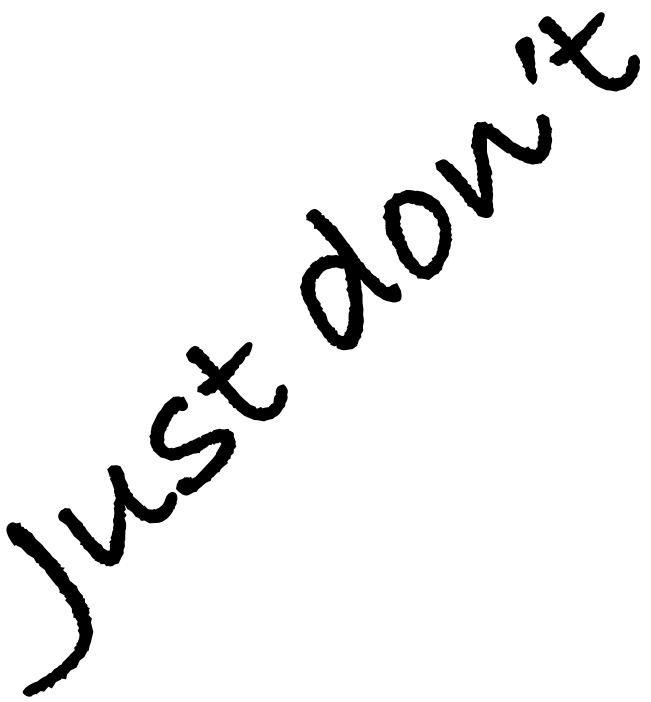
[Simmons, Nelson, Simonson: False-Positive Psychology, 2011]

Questionable Research Practices

Dos

- Limit researcher degrees of freedom.
- Stick to *a-priori* hypotheses, stopping rules, data clean-up, analysis plans.
- Be transparent: commit the plan in pre-registration.
- Publish the time-stamped pre-registration with study.
- Declare any deviation.

Don’ts



THE REPLICATION CRISIS

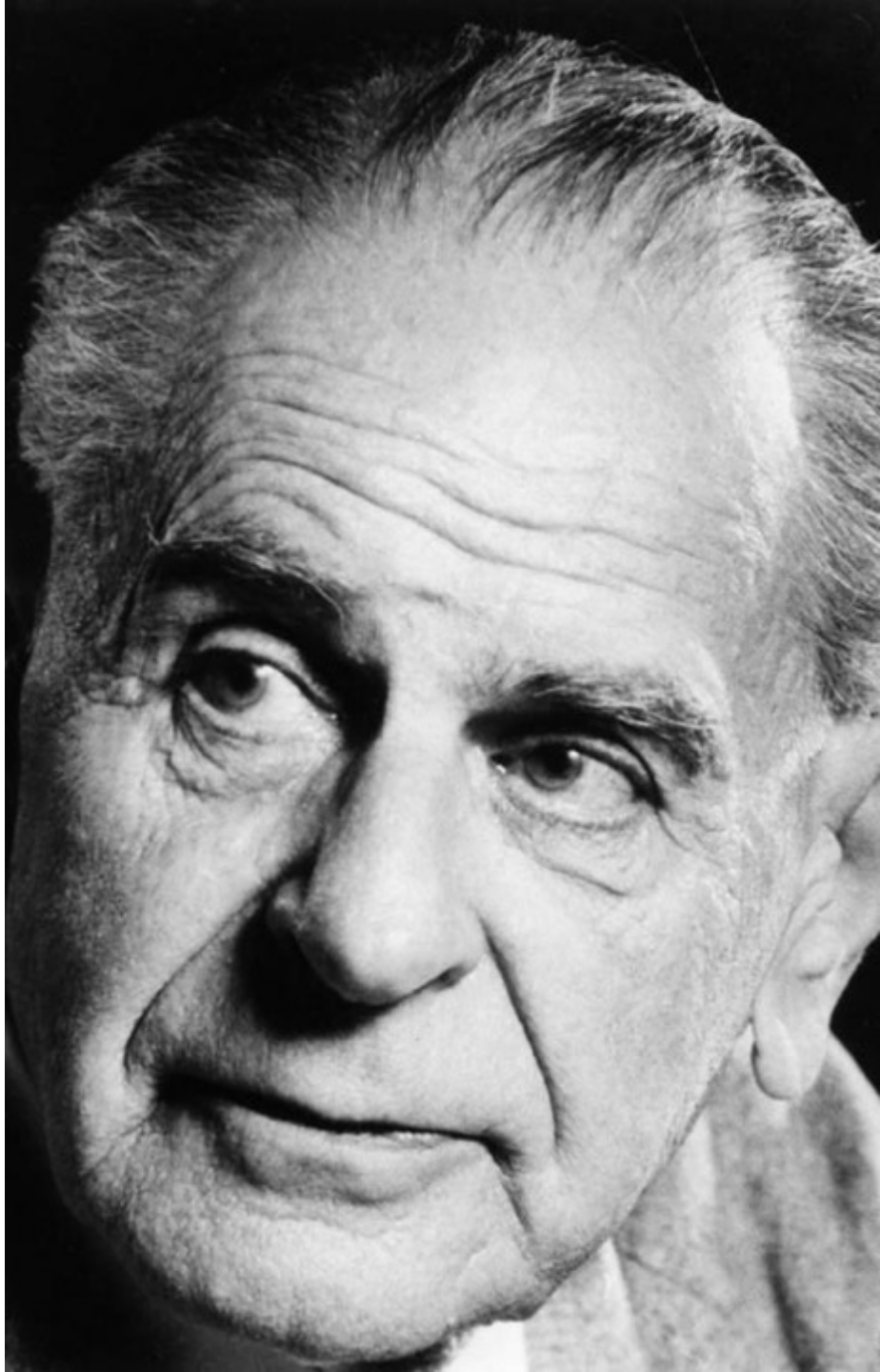
Learning Through Replications

“Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated ‘coincidence’“

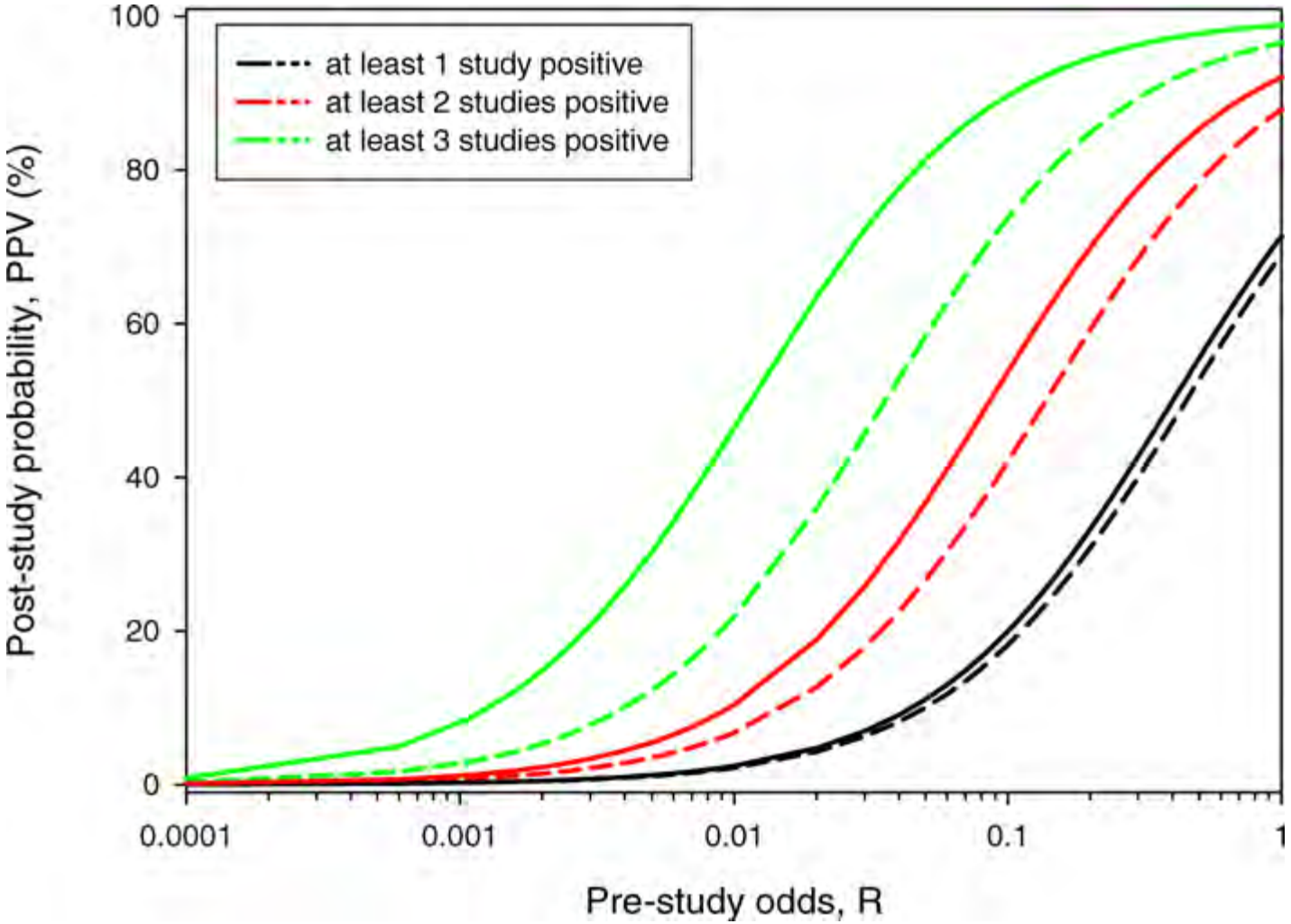
– Sir Karl Popper

Reproducibility is a defining feature of science.

A *replication crisis* is a state in science, in which scientists found that many findings cannot be replicated.

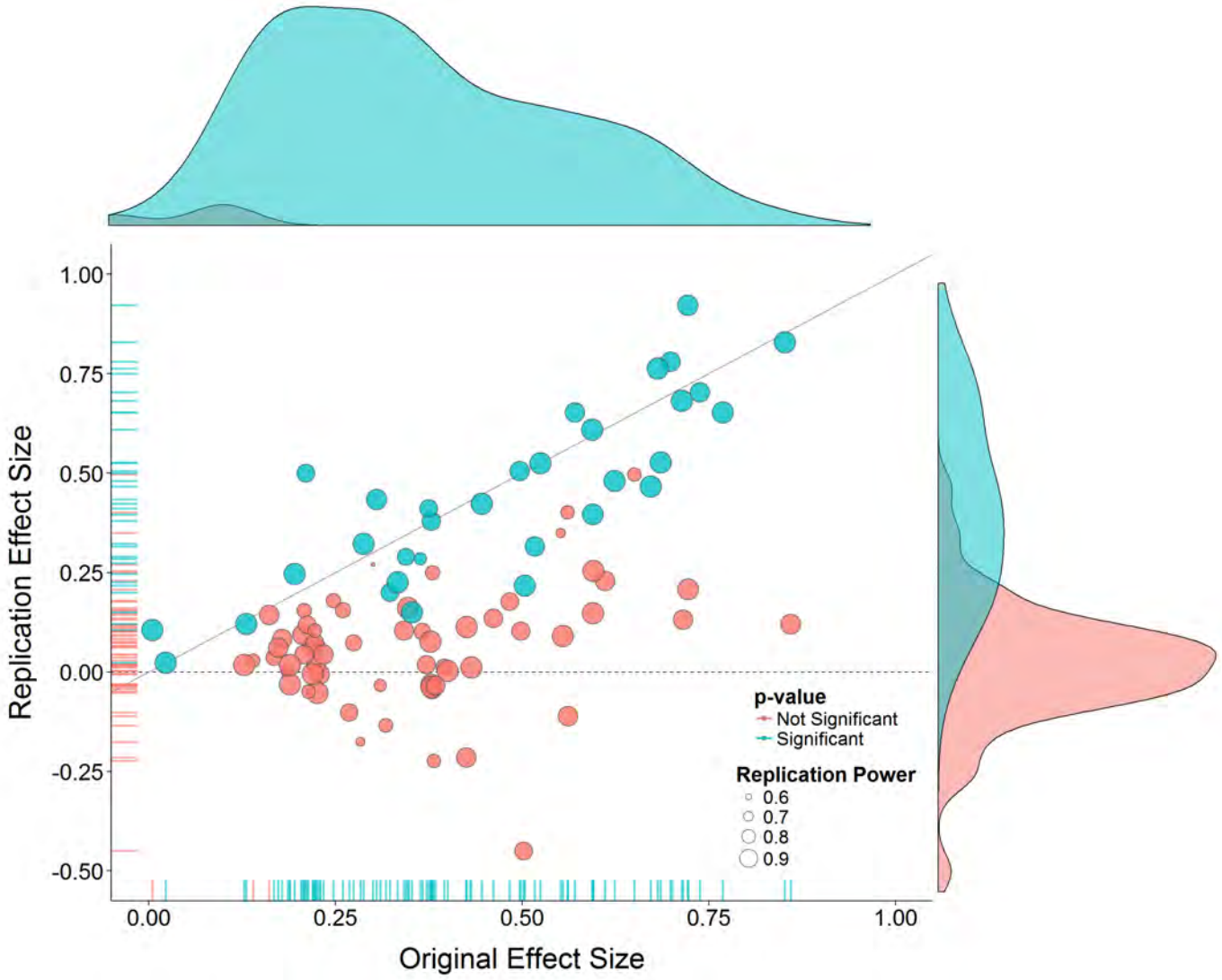


Good Replications Can Help



[Moonesinghe et al. 2007, Most Published Research Findings Are False—But a Little Replication Goes a Long Way]

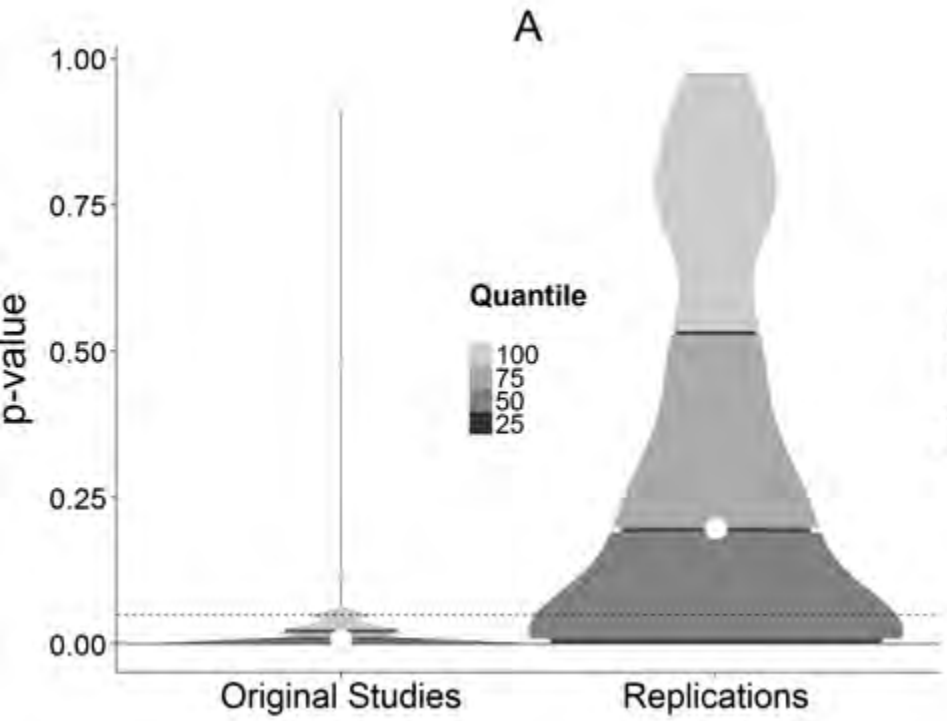
Replication Crisis in Psychology



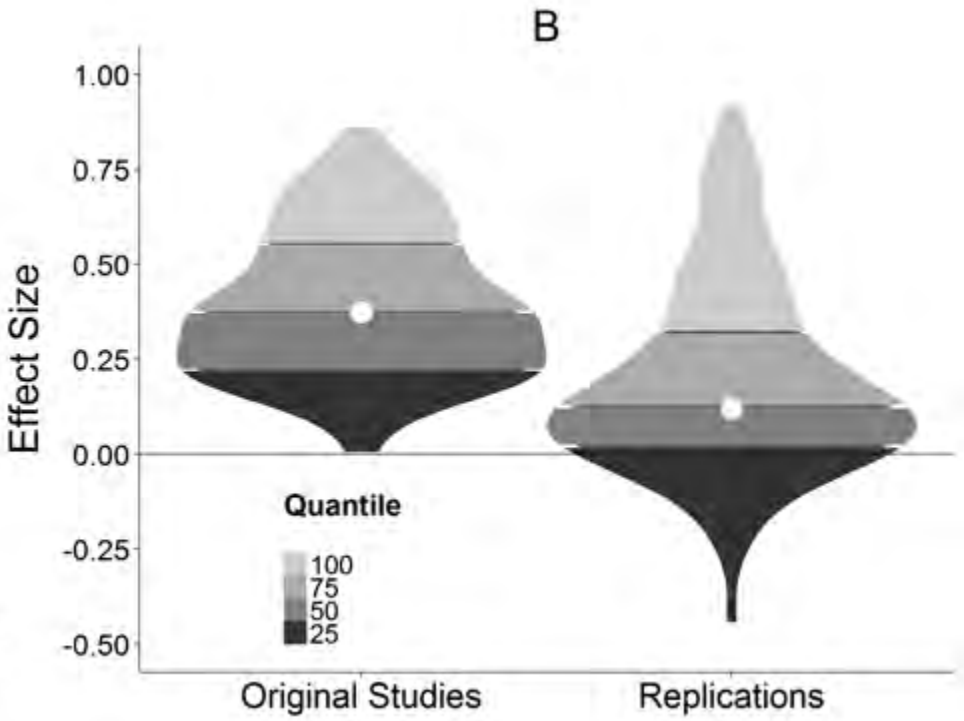
[The Reproducibility Project. Graphics courtesy of Fred Hasselman.]

Significance and Effect Size Estimates

Reported Significance



Reported Effect Sizes



[The Reproducibility Project. Graphics courtesy of Fred Hasselman.]

Few Replications in Cyber Security User Studies

In an SLR sample of **146** studies,

- **4** (**3%**) studies replicated original studies
- **2** (**1%**) of them were done by authors of the original study.
- **1** external replication was referenced.

- Replications seem few and far between.

- What kinds of replications did we find?
  - ***Direct Replication*:** As similar to the original as possible.
    Purpose: Check original finding (Reproducible?).
  - ***Conceptual Replication:*** Elements (intentionally) altered.
    Purpose: Expand range of theory's viable conditions.
- Only found conceptual replications, so far.

Implications
Replication Crisis

- Replications are few and far between in cyber security, yielding an exposure to a replication crisis.

- Even though some conferences encourage the submission of replications (e.g., SOUPS), few were published.

Learning Through Replications

Dos

- Do replications.
- Do research synthesis.
- Replicate upstream tools (instruments etc.)
- Prepare for downstream replication by others:
  - Pre-registration
  - Study protocol
  - Study materials
  - Datasets
  - Analysis scripts

Don’ts

- Do not take unreplicated results at face-value.
- Do not give in into the fear of embarrassment.

Begley's
Six Red Flags for Suspect Work

1. Were the experiments performed blinded?
2. Were the basic experiments repeated?
3. Were all results presented?
4. Were there positive and negative controls?
5. Were the [instruments] validated?
6. Were the statistical tests appropriate?

[Begley, 2013 – Six red flags for suspect work]

Conclusions

- Evidence-based methods have great potential, **but** our field's implementation is often flawed.
- Design for falsification & reproducibility is key.
- Empirical analysis of state-of-play highlights issues in individual studies as well as the field.
- Imprudent design choices and biases cause a low posterior probability of positive reports.
- Replications crucial, yet virtually non-existent.

Acknowledgment

This work was conducted as part of the EPSRC/NCSC Research Institute in the Science of Cyber Security II.

It was in parts funded by the RISCS/NCSC grant “Pathways to Evidence-Based Methods in Cyber Security,” especially Pathway I: Strength of Evidence and Meta-Analysis.

The investigator is being supported by the European Research Council (ERC) Starting Grant “Confidentiality-Preserving Security Assurance (CASCAde),” GA nº 716980.