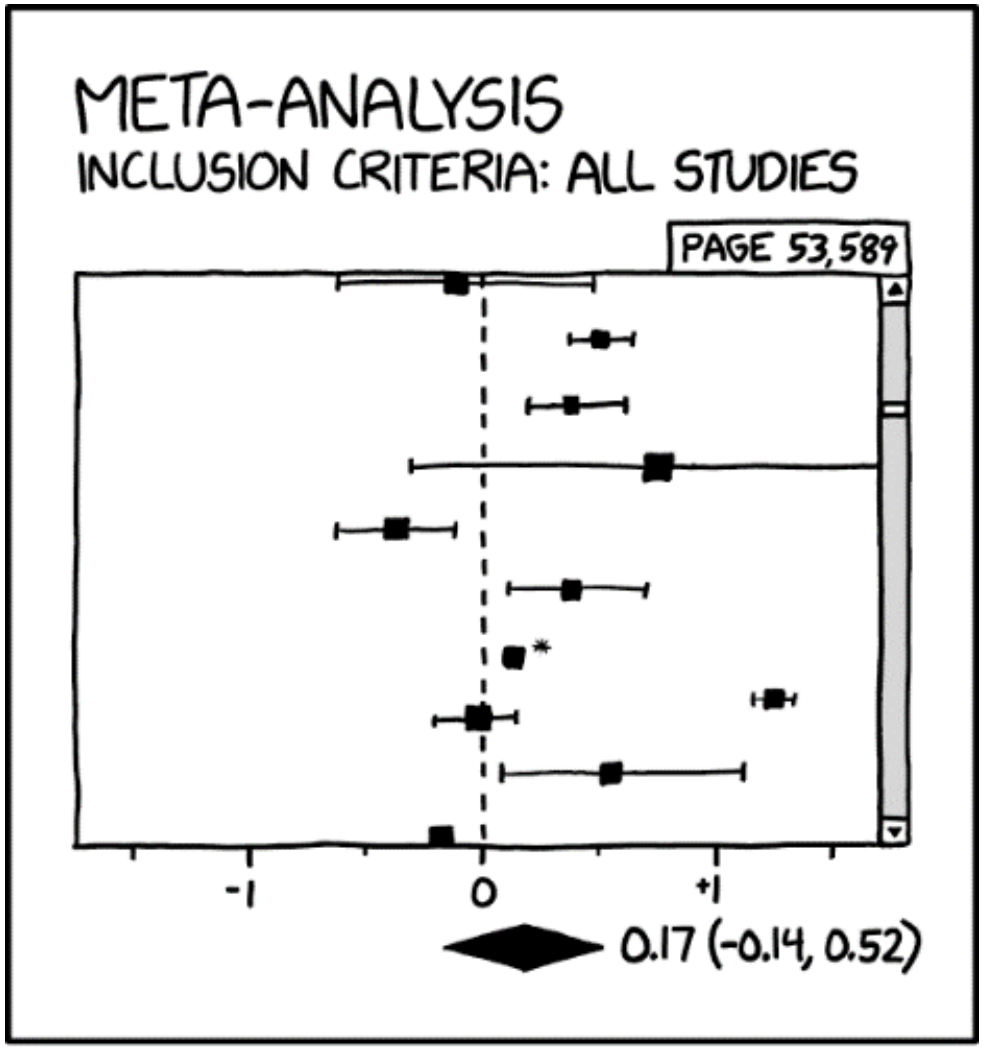


BAD NEWS: THEY FINALLY DID A META-ANALYSIS OF ALL OF SCIENCE, AND IT TURNS OUT IT'S NOT SIGNIFICANT.

https://xkcd.com/2755/

I have the (Developer) Power!
Supporting Power Analysis and Statistical Reporting in Usable Security and Privacy

Anna-Marie Ortloff[1], Christian Tiefenau[2], Matthew Smith[1,2]
[1]: University of Bonn, [2]: Fraunhofer FKIE

A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes

Anna-Marie Ortloff
University of Bonn
Bonn, Germany
ortloff@cs.uni-bonn.de

Julia Angelika Grohs
University of Bonn
Bonn, Germany
s6jugroh@uni-bonn.de

Simon Lenau
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
lenau@cispa.de

Matthew Smith
University of Bonn
Bonn, Germany
Fraunhofer FKIE
Bonn, Germany
smith@cs.uni-bonn.de

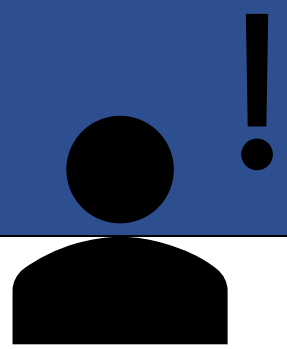
S'23

CHI'25

Behavioural Security Group

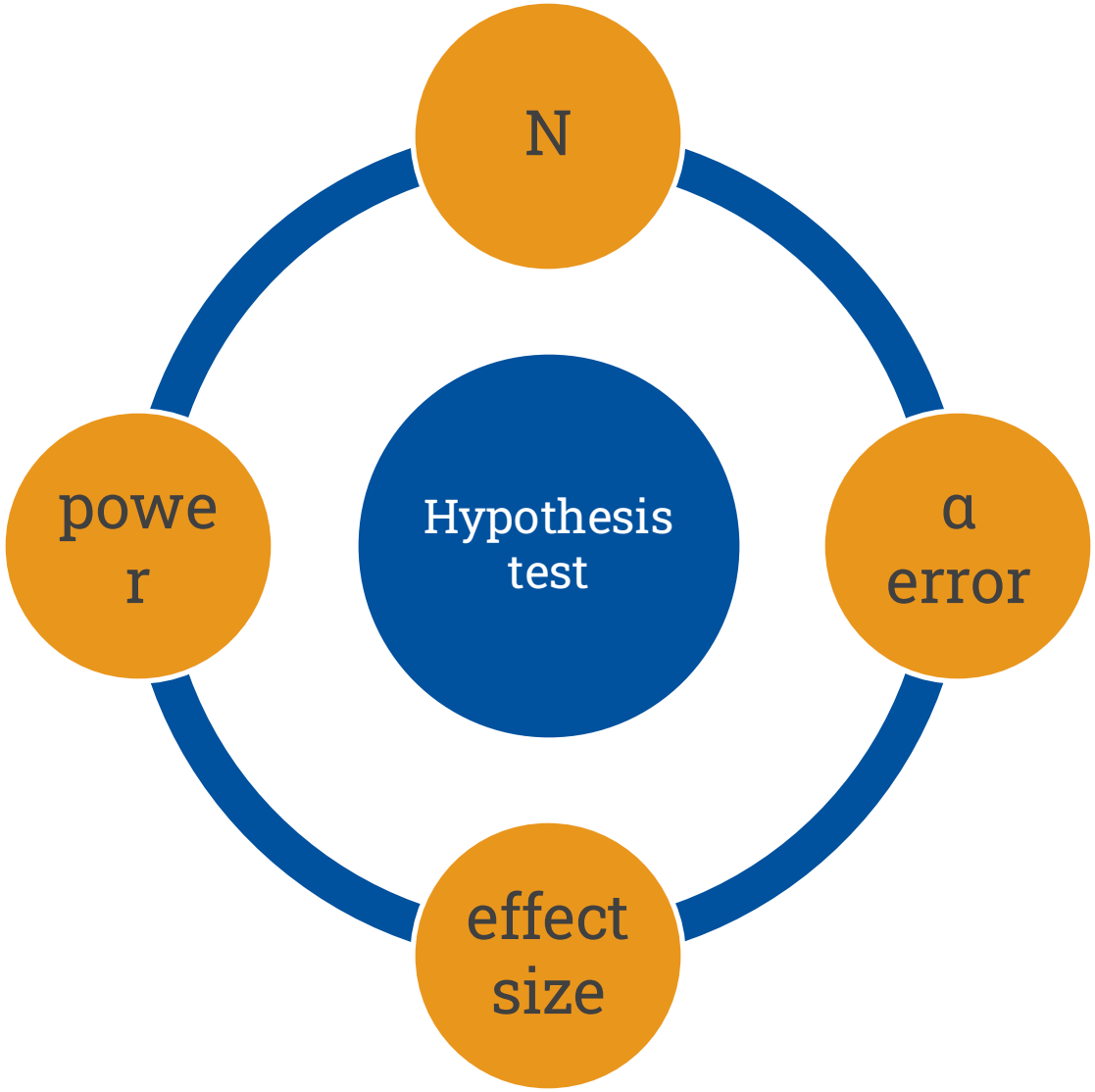
UNIVERSITÄT BONN

[Find recommendations here]

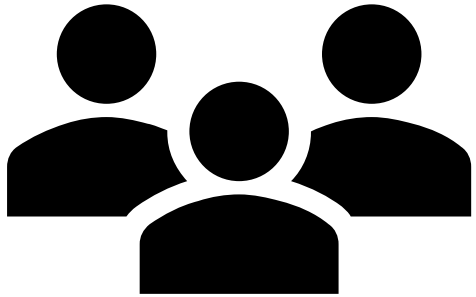


Background

Hypothesis Testing



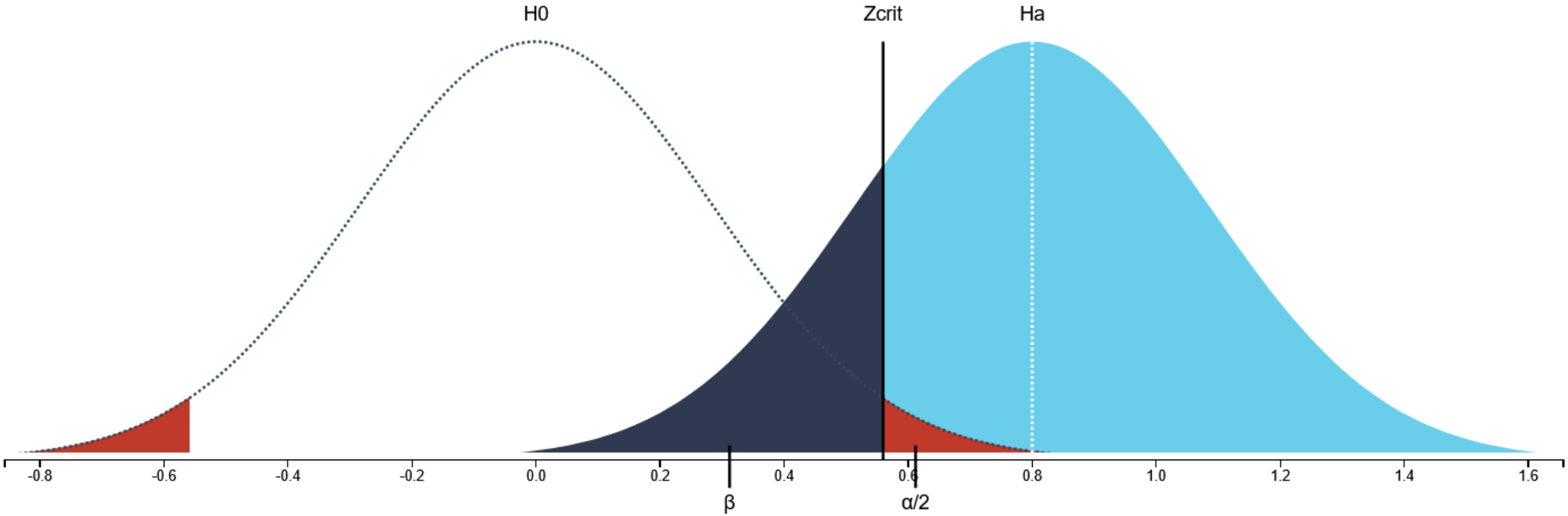
Sample Size N



Statistical power

The probability of detecting an effect, if a true effect exists.

Alpha Error

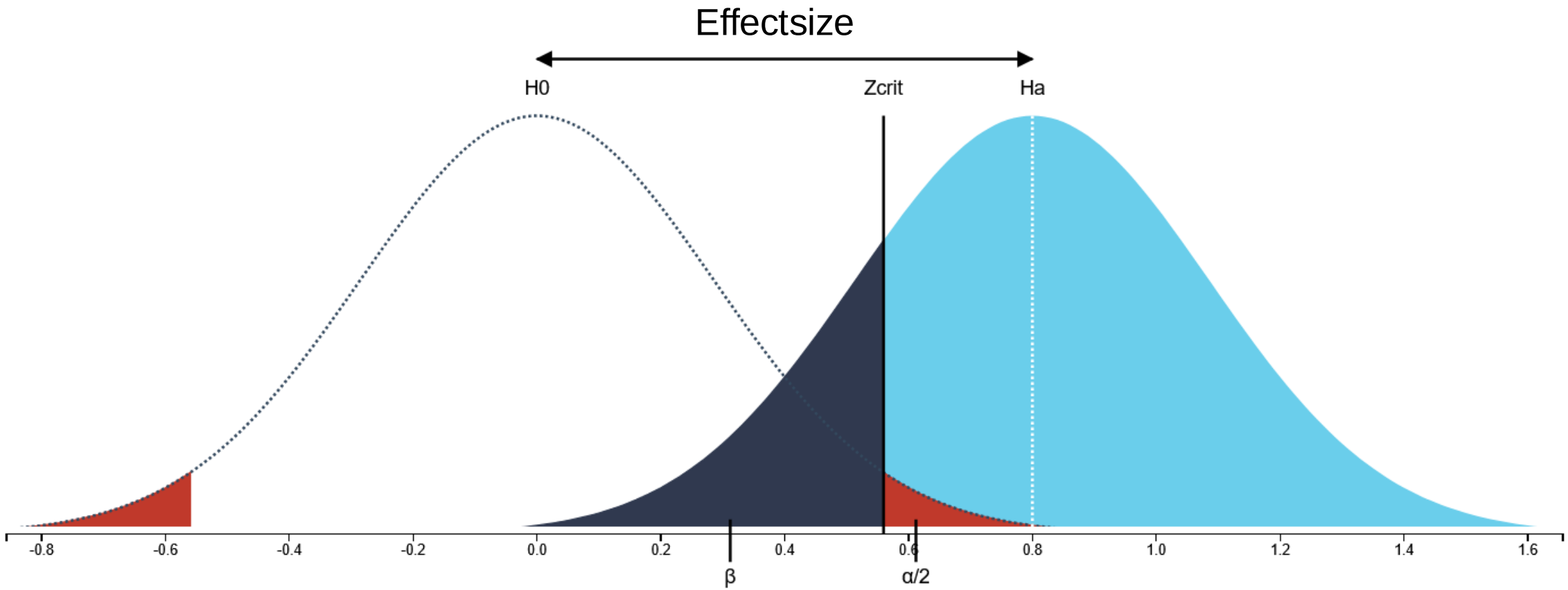


https://rpsychologist.com/d3/nhst/

Effect size

The strength of the relationship
of predictor variables with outcome variables.

Effect size



https://rpsychologist.com/d3/nhst/

Example Effect Size Measures

Odds ratio

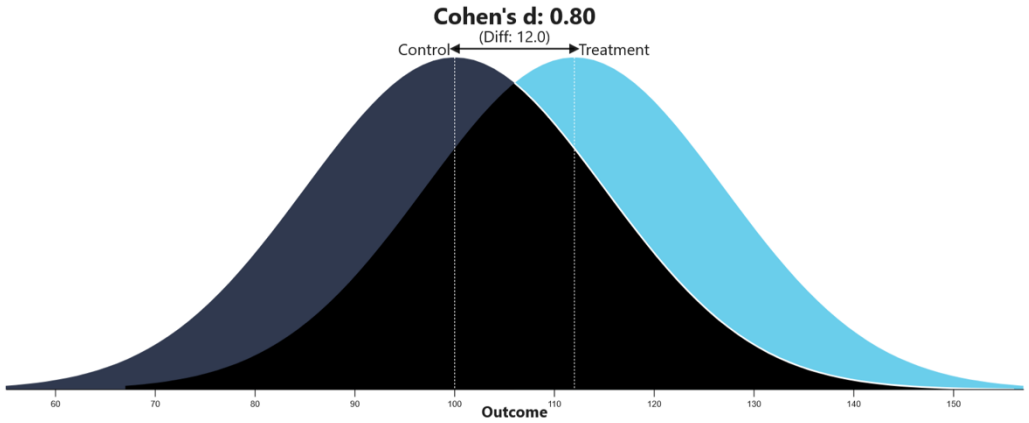
- Ratio of two odds

	Adopted PWM	Did not adopt PWM
Group 1	10	30
Group 2	20	20

- OR=1 means the odds of an outcome are the same in both groups.

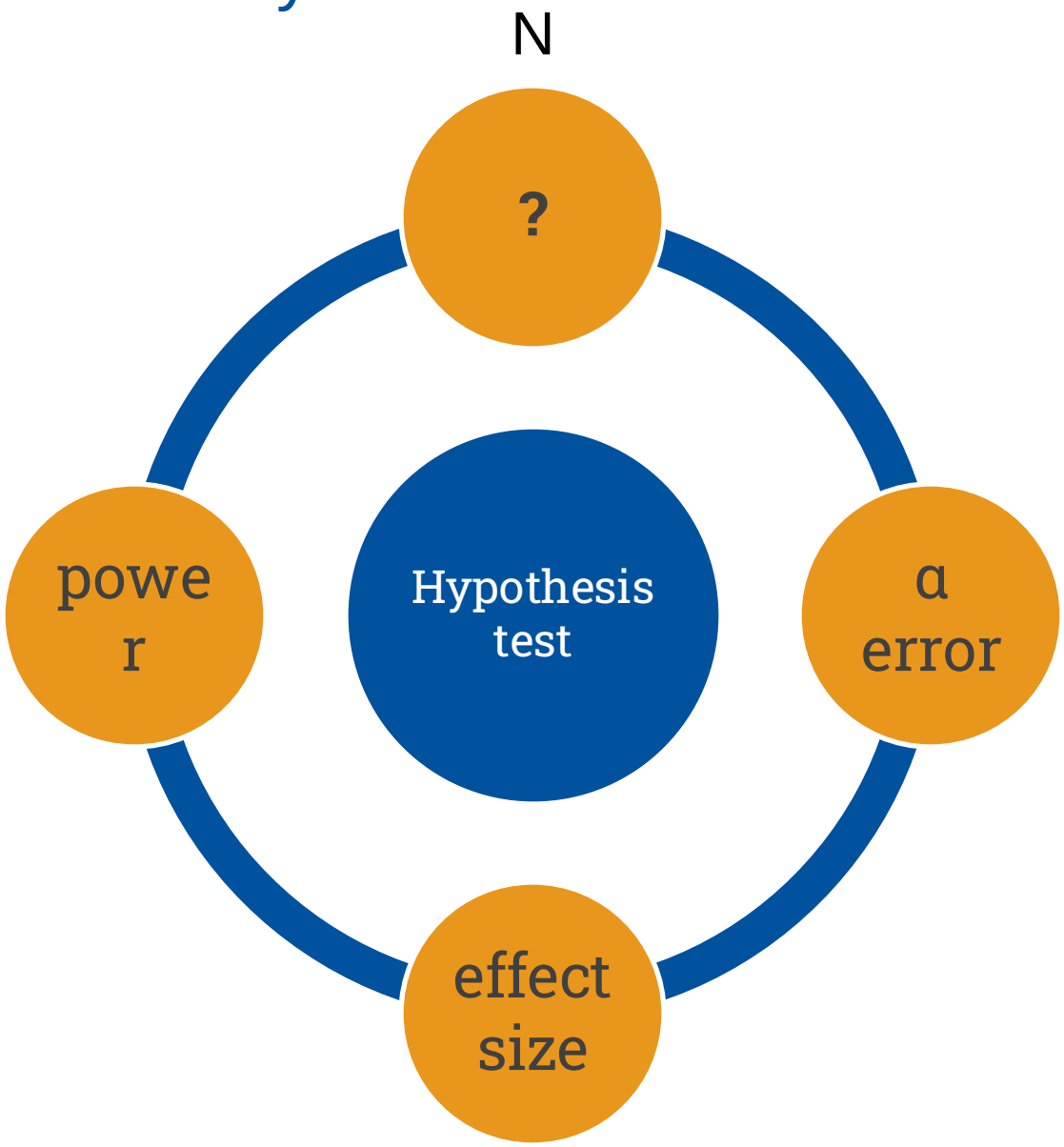
Cohen‘s d

- Normalized difference in means
- d = 0.8 is difference of 0.8 standard deviations between the means

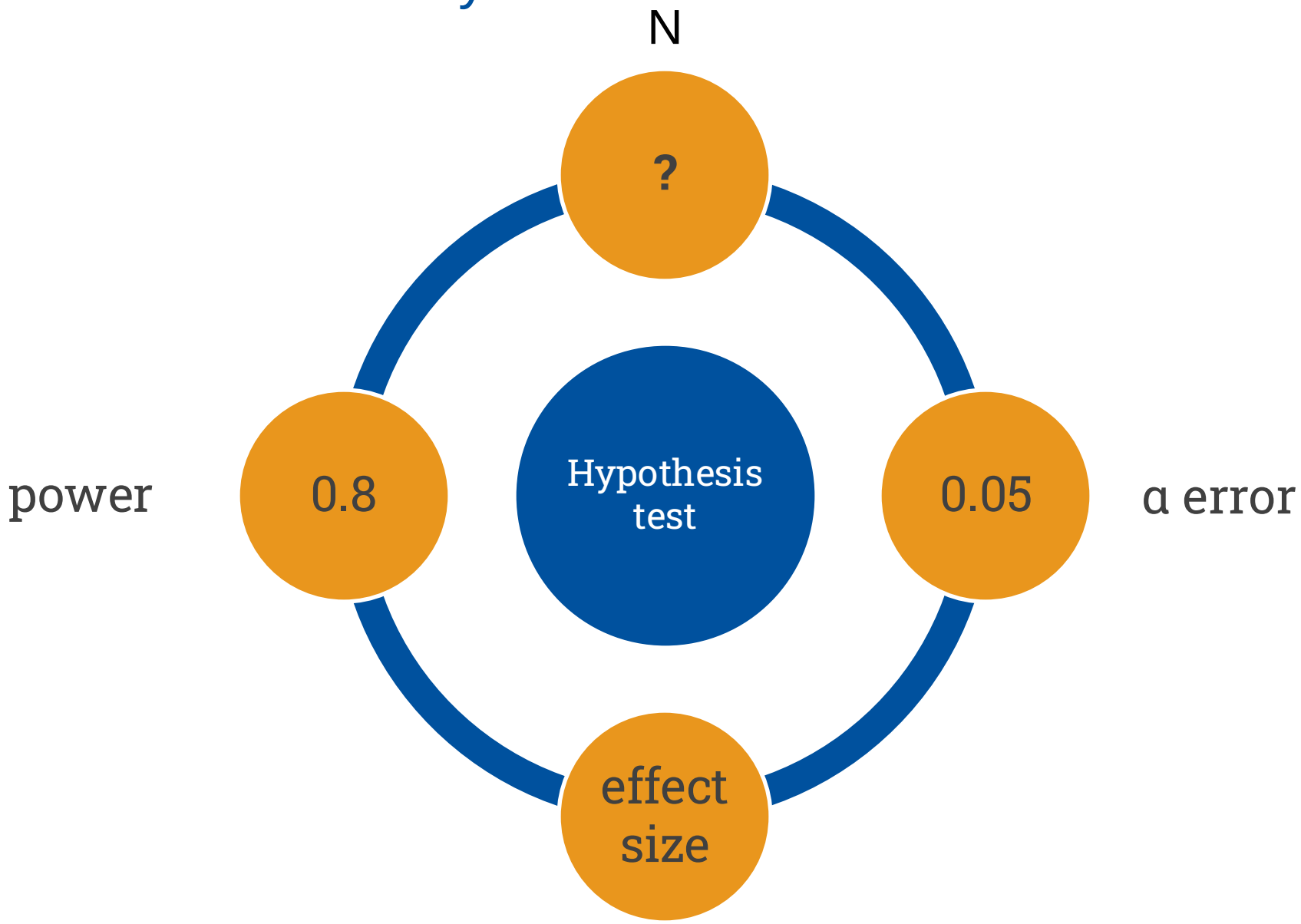


https://rpsychologist.com/cohend/

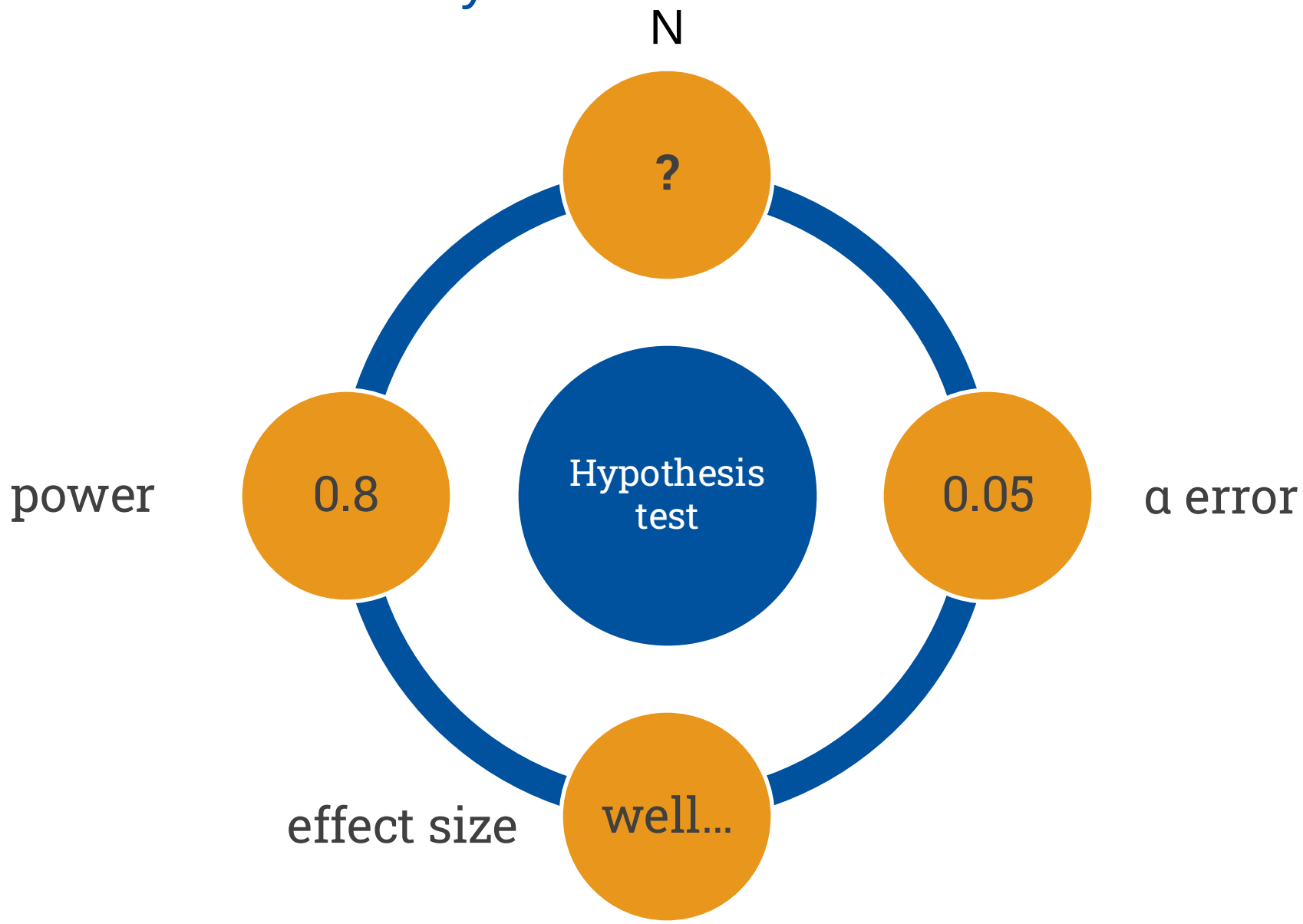
A Priori Power Analysis



A Priori Power Analysis



A Priori Power Analysis

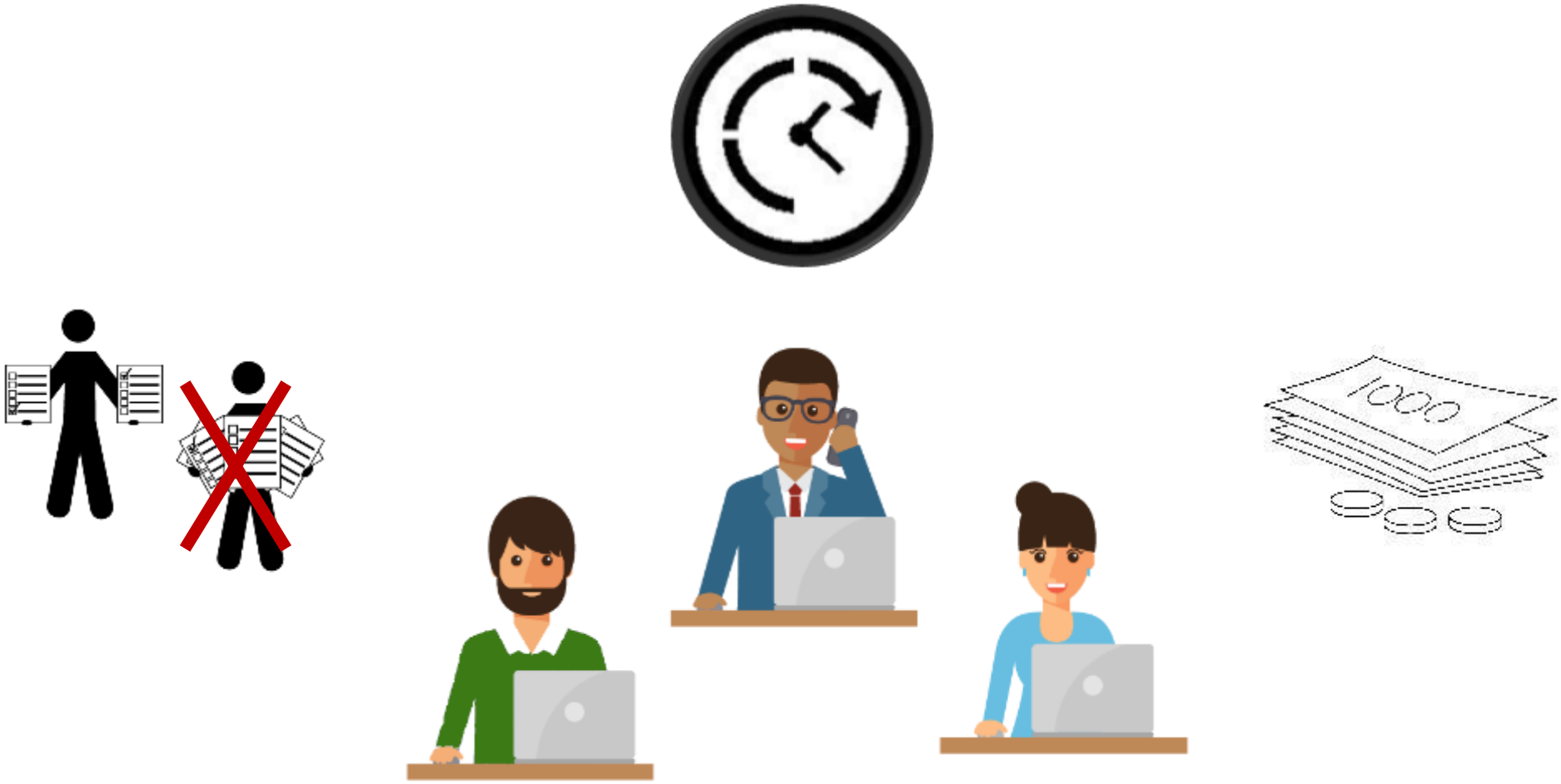


Behavioural Security Group

UNIVERSITÄT BONN

Do we have the power?

Developer-Centered Usable Security (DCUS)



Method

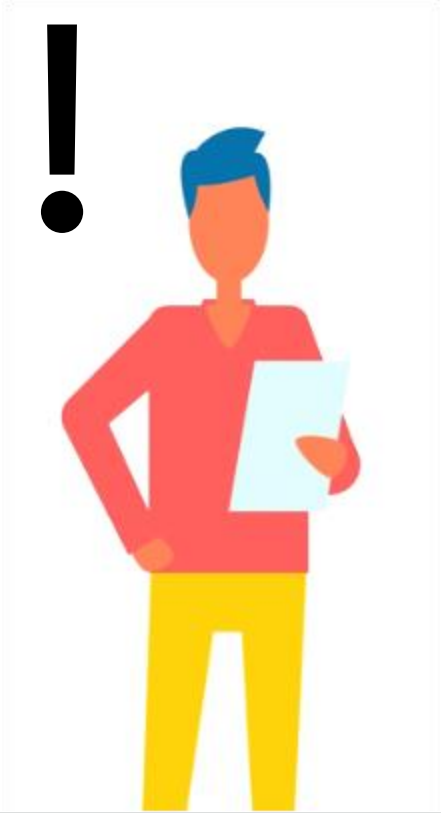
Literature Collection

- SOUPS, USENIX Security, S&P, CCS, ICSE, USP Tracks of CHI
- 2010 - 2021
- Include user study
- Participants: software developers, similar expert users, or proxies
- Domain of usable security and privacy

- 54 papers
- including 64 studies, 467 hypothesis tests, 413 variables

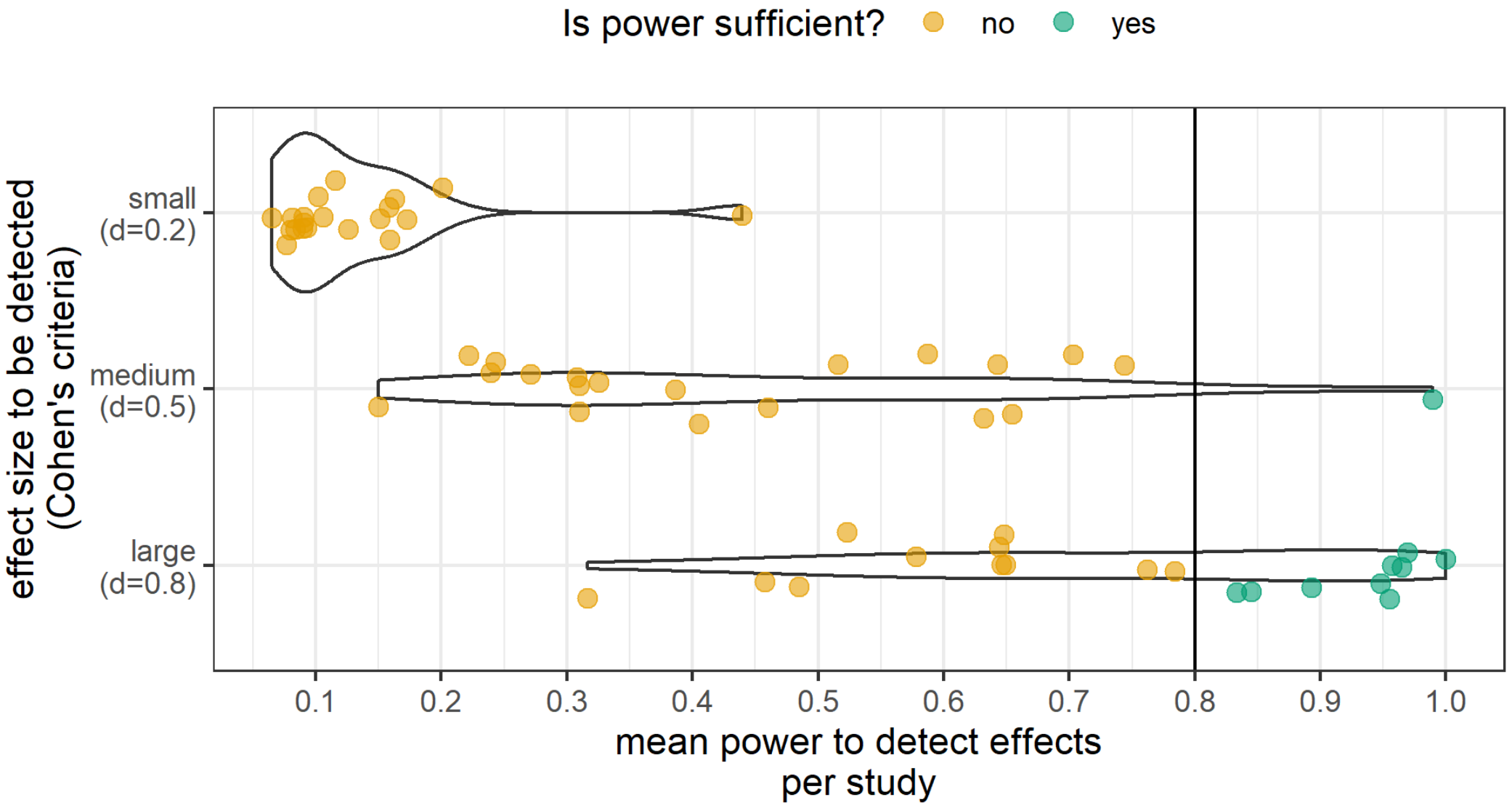
Data Structure

- Relevant information on power and effect sizes in these studies



Power Meta-Analysis (simulated a-priori power analysis)

Power Meta-Analysis (simulated a-priori power analysis)



We‘re not using power analysis.

In SOUPS and CHI USP publications from 2020/2021

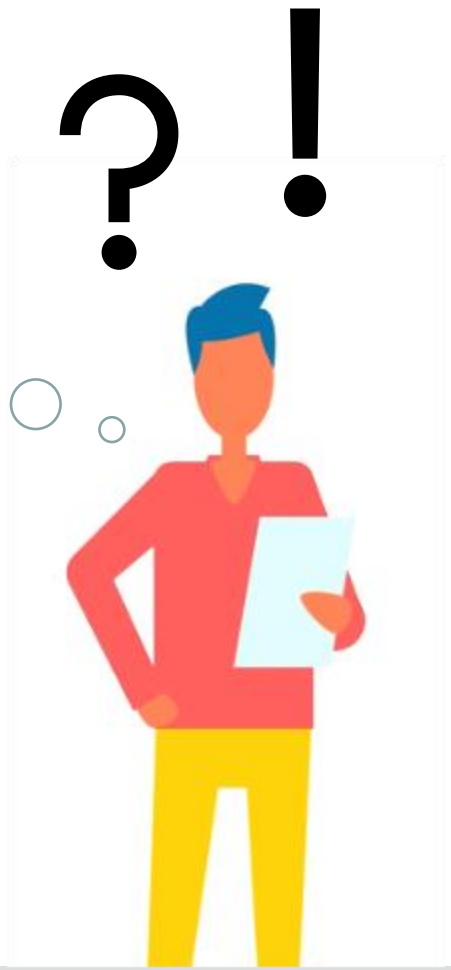
only 8 of 74 (10.8%) quantitative papers used a priori power analysis

Behavioural Security Group

UNIVERSITÄT BONN

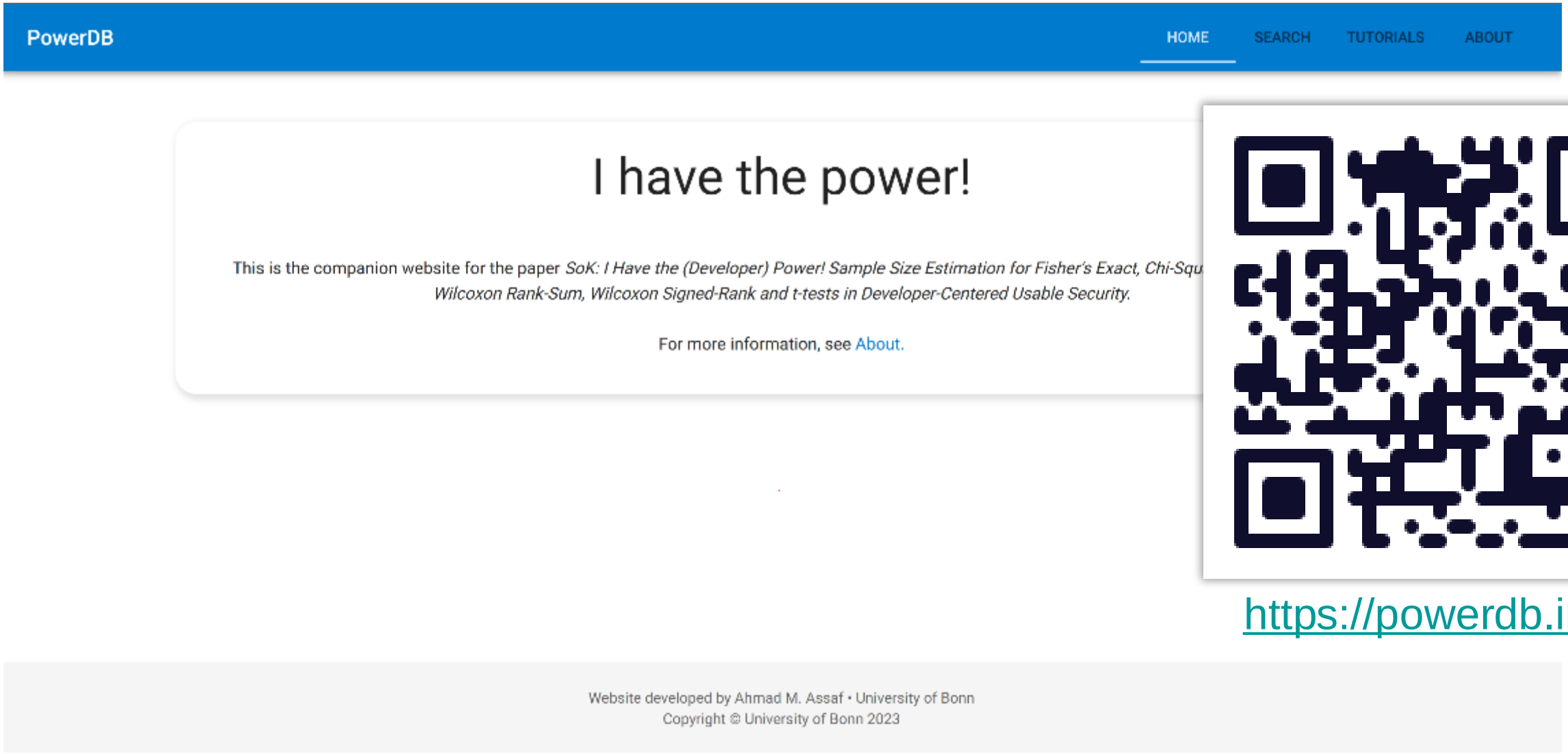
How can we get the power?

Which effect size should I use?



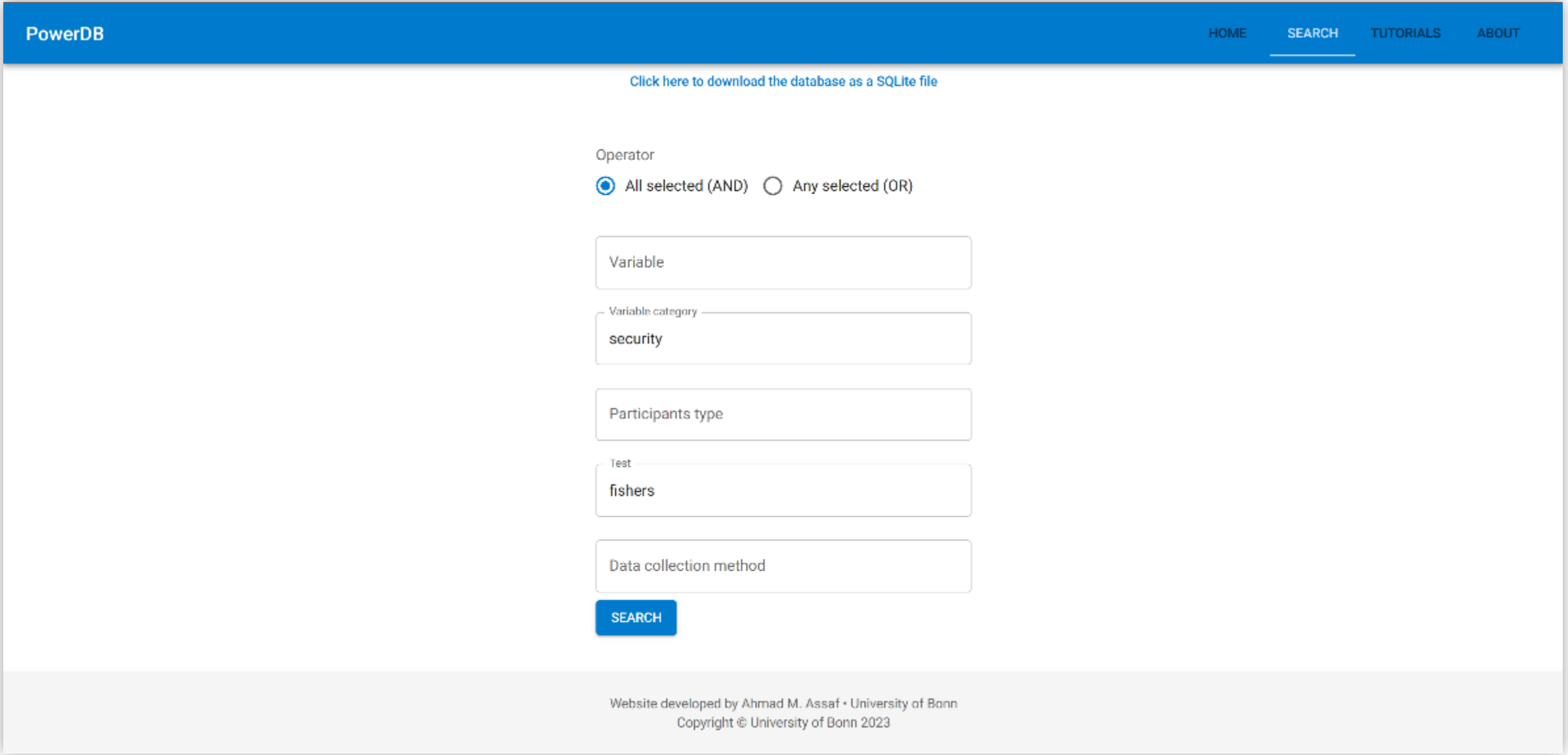
- Use general guidelines for large, medium, small effects
- Use context specific guidelines for large, medium, small effects
- Literature research
- Do a pilot study
- Decide on the smallest effect size of interest

Database + Companion Website

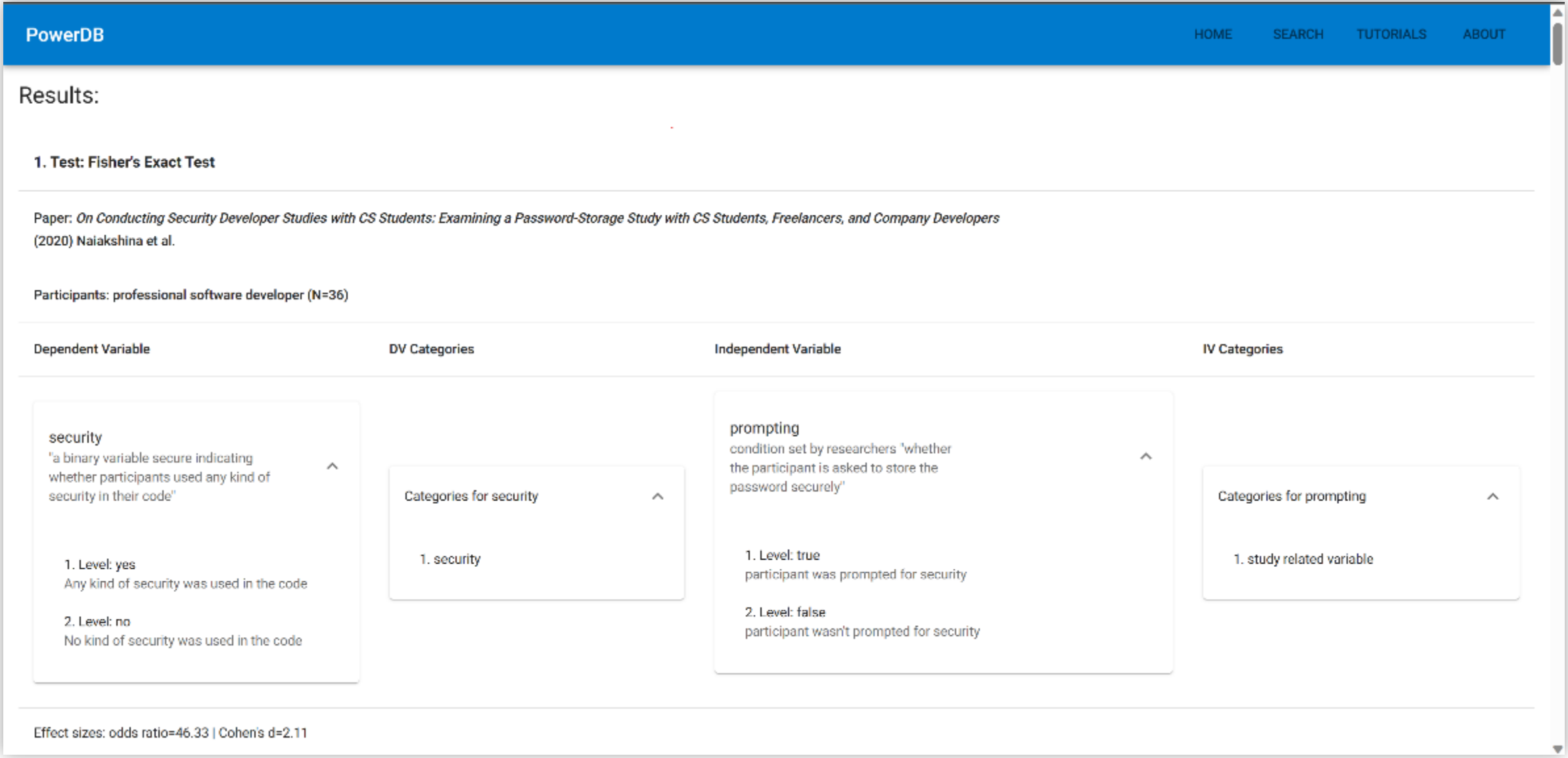


https://powerdb.info/

Searching the database



Searching the database



PowerDB

HOME SEARCH TUTORIALS ABOUT

Introduction

- Introduction
- Fisher's Exact Test
- Chi-Squared Test
- McNemar's Test
- Independent t-Test
- Paired t-Test
- Wilcoxon Rank-Sum Test
- Wilcoxon Signed-Rank Test

A Guide to Power Analysis for Hypothesis Tests with One Categorical Independent Variable with Two Groups

Anna-Marie Ortloff *University of Bonn*
Christian Tiefenau *University of Bonn*
Matthew Smith *University of Bonn, Fraunhofer FKIE*

What can I expect?

What is Power Analysis?

Four parameters are relevant to power analysis: power, the significance criterion (i.e. the \(\alpha\) error level), the reliability of the sample results or sensitivity of the test, and the effect size [2]. These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. Such calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three, although it is also possible to determine both \(\alpha\) and power if a ratio for \(\alpha\) and \(\beta\) is given together with the other two parameters - this is termed compromise power analysis [5]. The other four flavors are summarized, e.g. by Cohen [2] in Chapter 1.5.

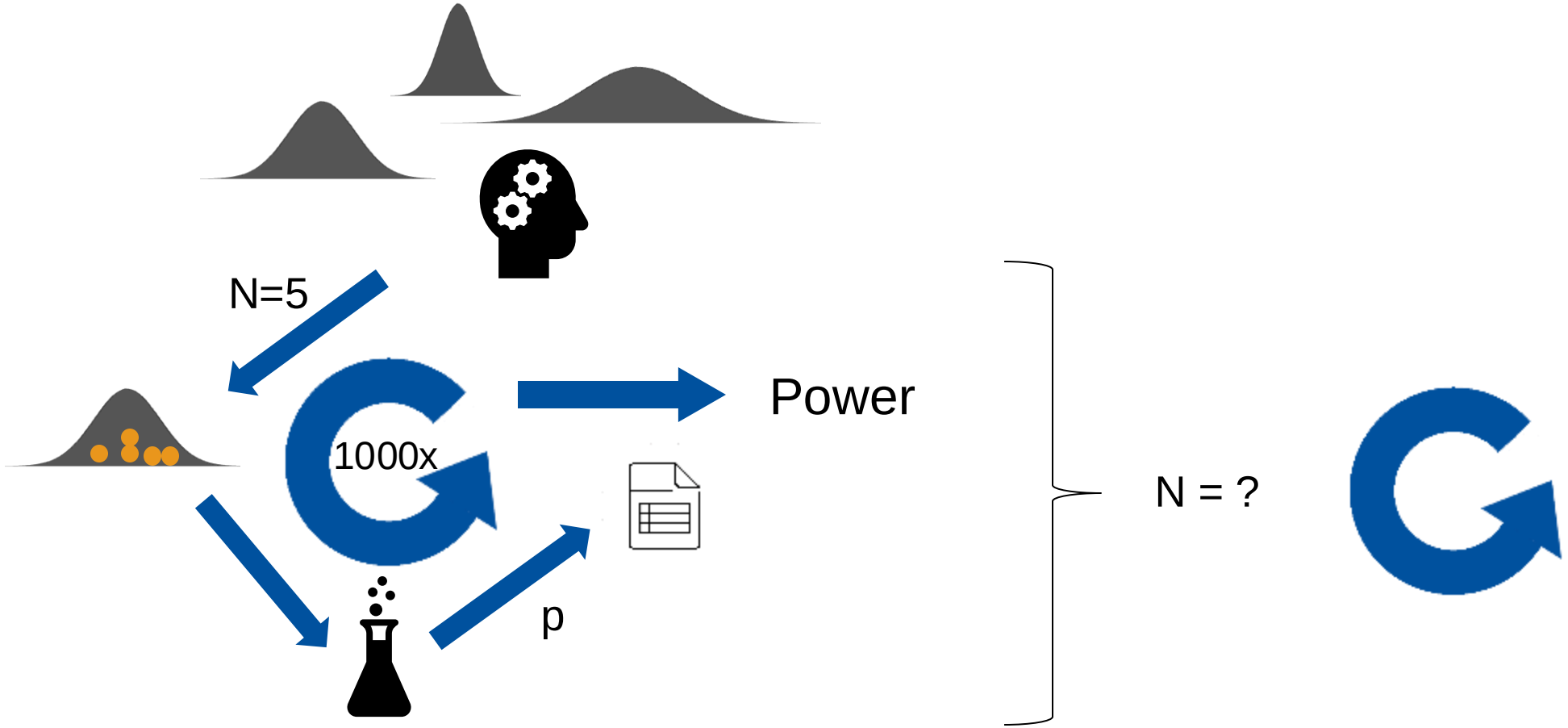
The Four Parameters Explained

All of the mentioned parameters are explained here. Click the cards below to see the definition.

- **Power**
- **Significance Criterion**
- **Reliability (Sample Size)**
- **Effect Size**

Example Power Analysis

Simulation for Power Analysis



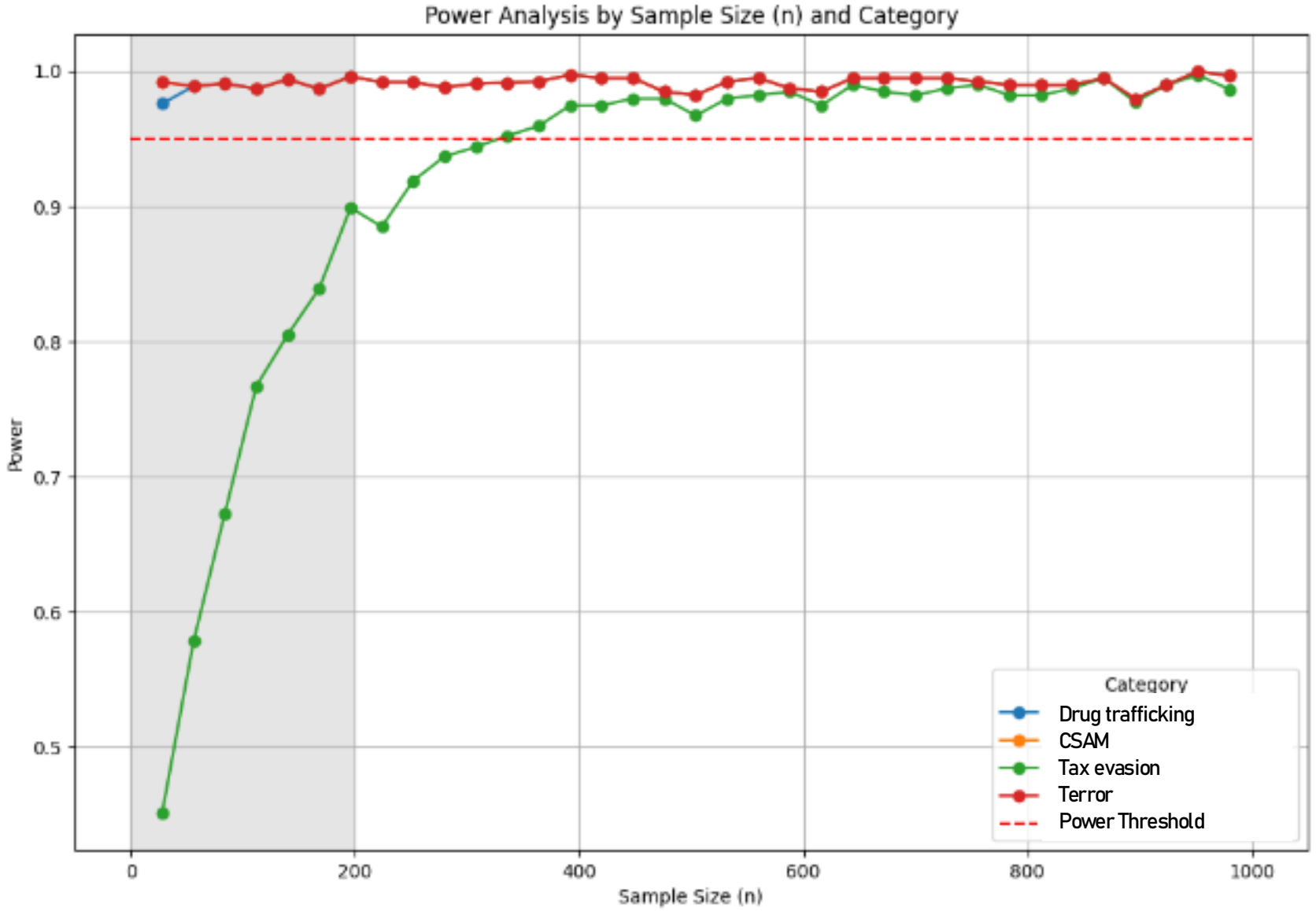
Example 1

"Not the Right Question?" A Study on Attitudes Toward Client-Side Scanning with Security and Privacy Researchers and a U.S. Population Sample

Lisa Geierhaas*, Florin Martius*, Arthi Arumugam†, Matthew Smith‡
*University of Bonn, {geierhaa,martius}@cs.uni-bonn.de
†University of Bonn, arumugam@uni-bonn.de
‡University of Bonn, Fraunhofer FKIE, smith@cs.uni-bonn.de

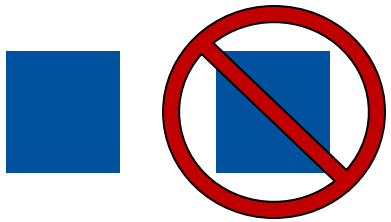
S&P'25

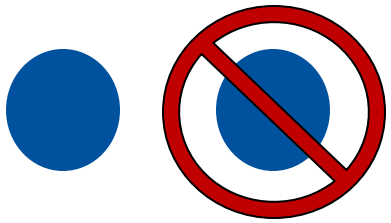
Results from Simulation

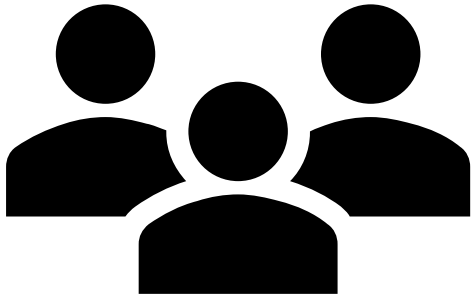


Example 2

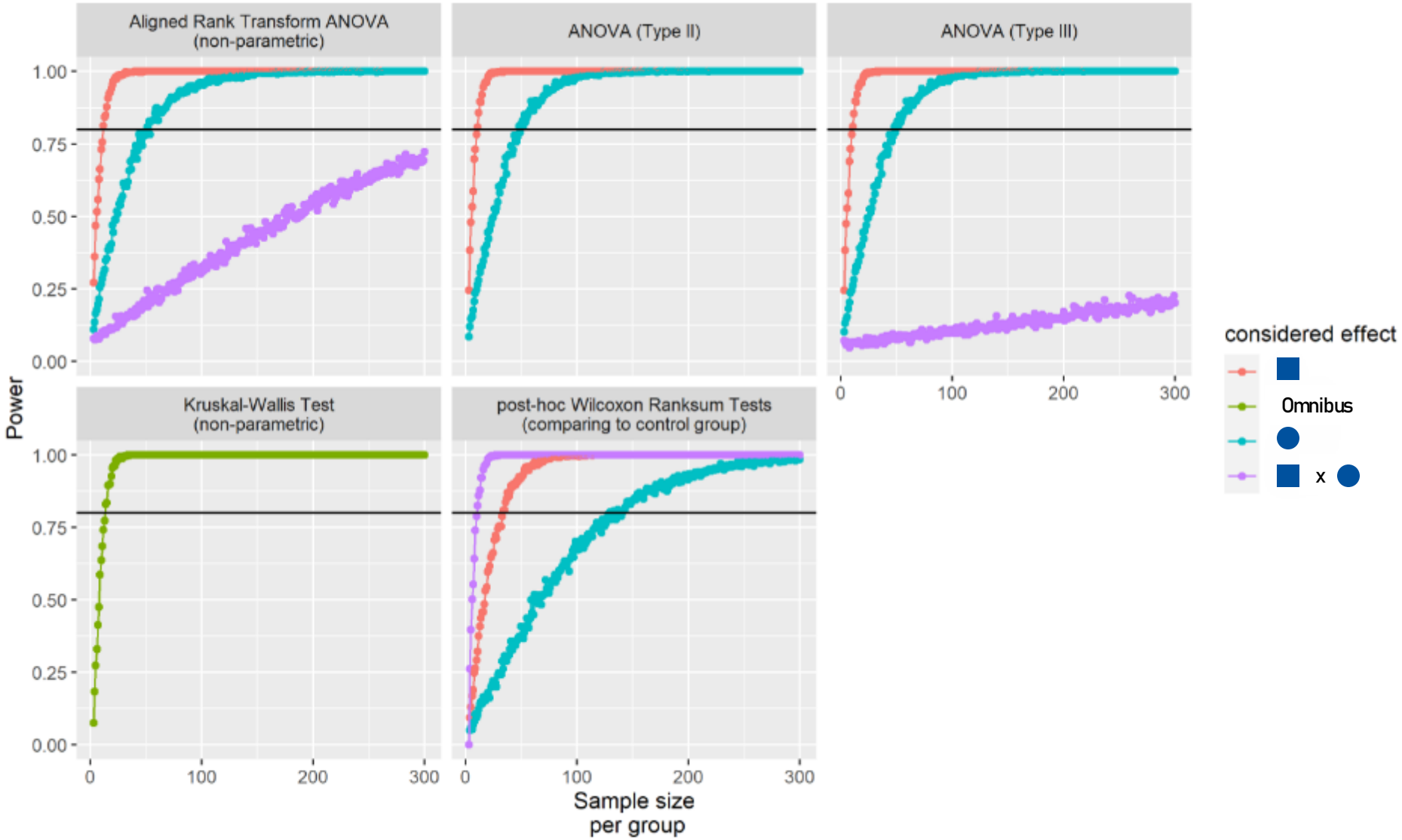
- Currently unpublished work
  - Comparing two types of intervention in a fully crossed design
- Effect size estimate based on
  - small-N pilot
  - prior work with raw data/frequencies published in the paper







Results from Simulation



Approach Comparison

Simulation

- One way: 14 / group
- Multi-way:
  - ● 48 / group
  - ■ 11 / group

- Post-hoc comparisons
  - ● 130 / group
  - ■ 34 / group
  - ●+■ 11 / group

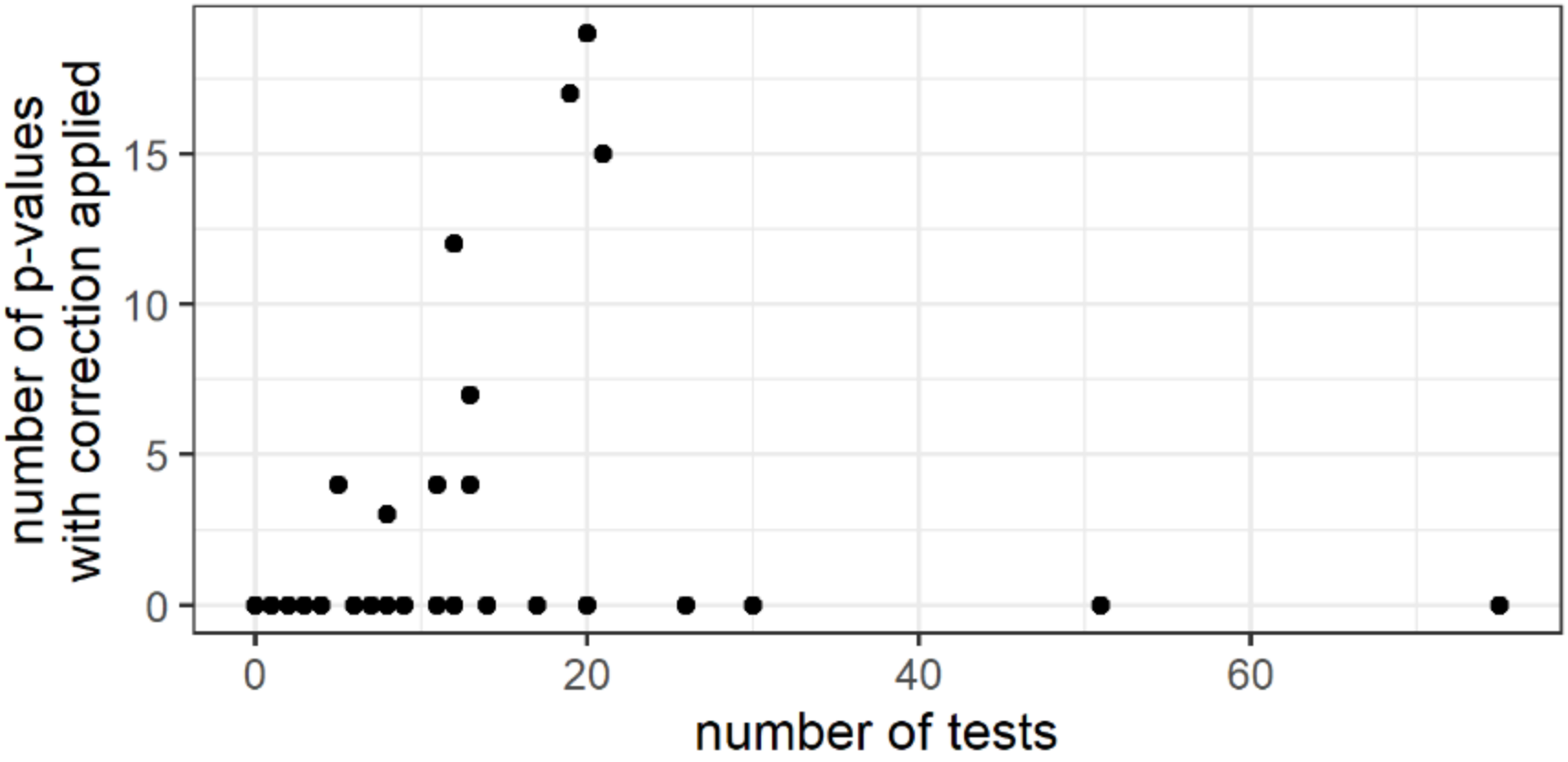
Direct

- One-way: 15 / group
- Multi-way
  - ● 55 / group
  - ■ 14 / group

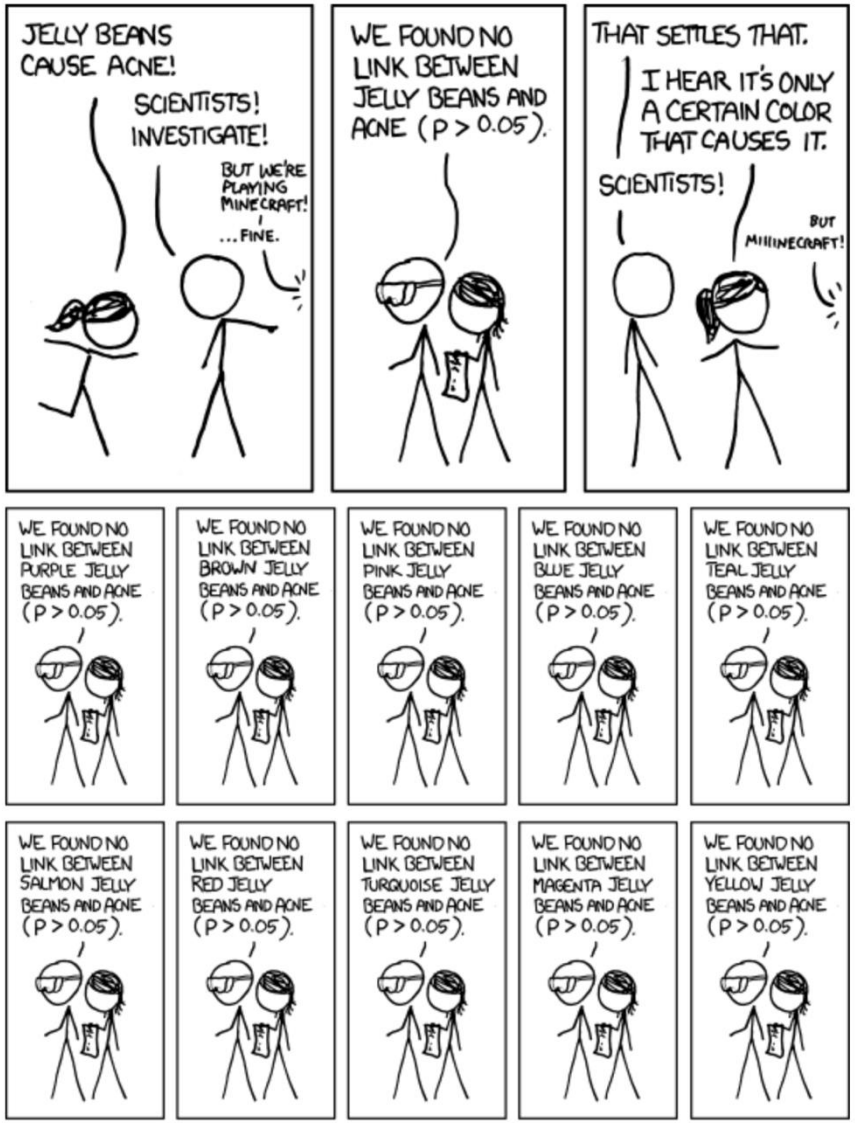
- Post-hoc comparisons:
  - ● 159 / group
  - ■ 29 / group
  - ●+■ 9 / group

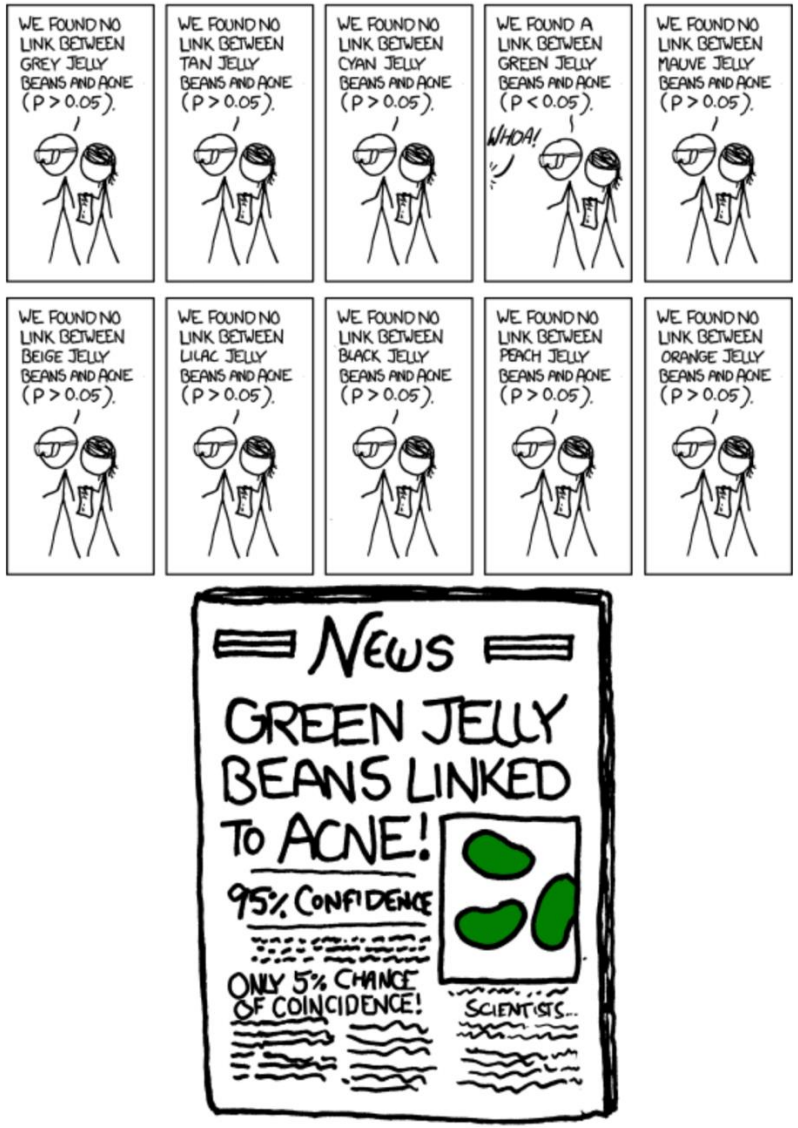
More Findings from DCUS

Many Statistical Tests



Too? Many Statistical Tests

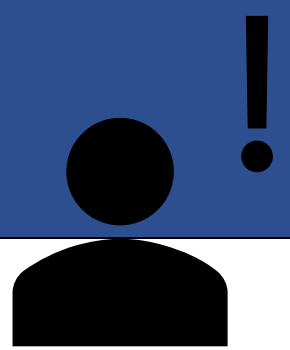




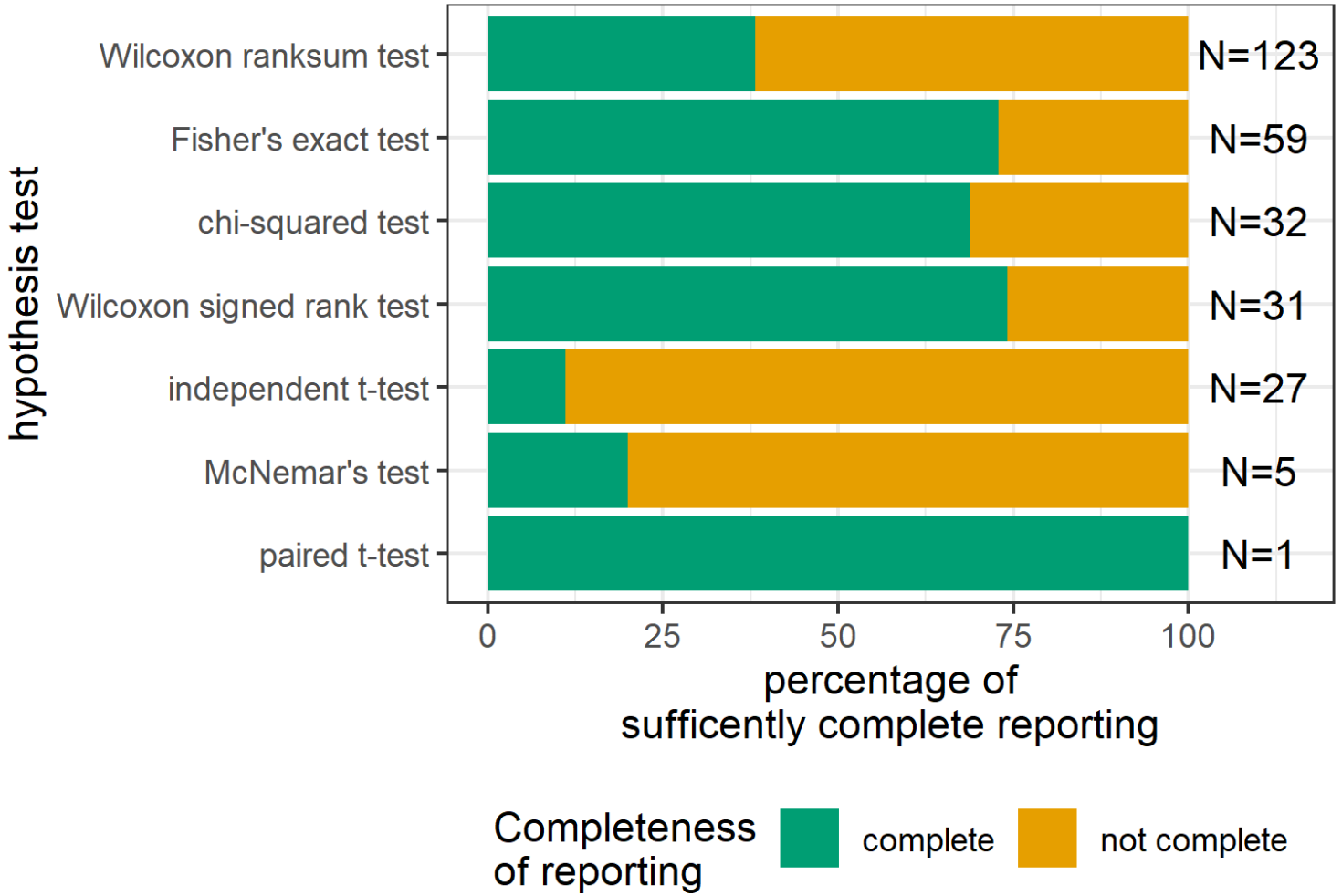
https://xkcd.com/882/

Recommendation

Specify if you are using
confirmatory or exploratory analysis methods.

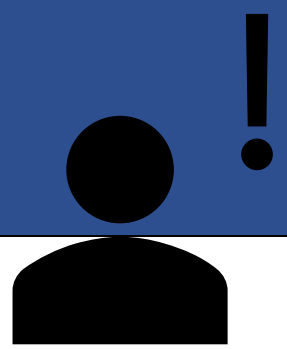


Problems with Reporting



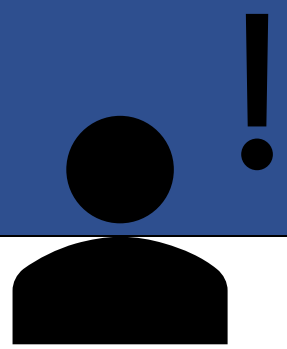
Recommendation

Report complete and appropriate descriptive statistics.

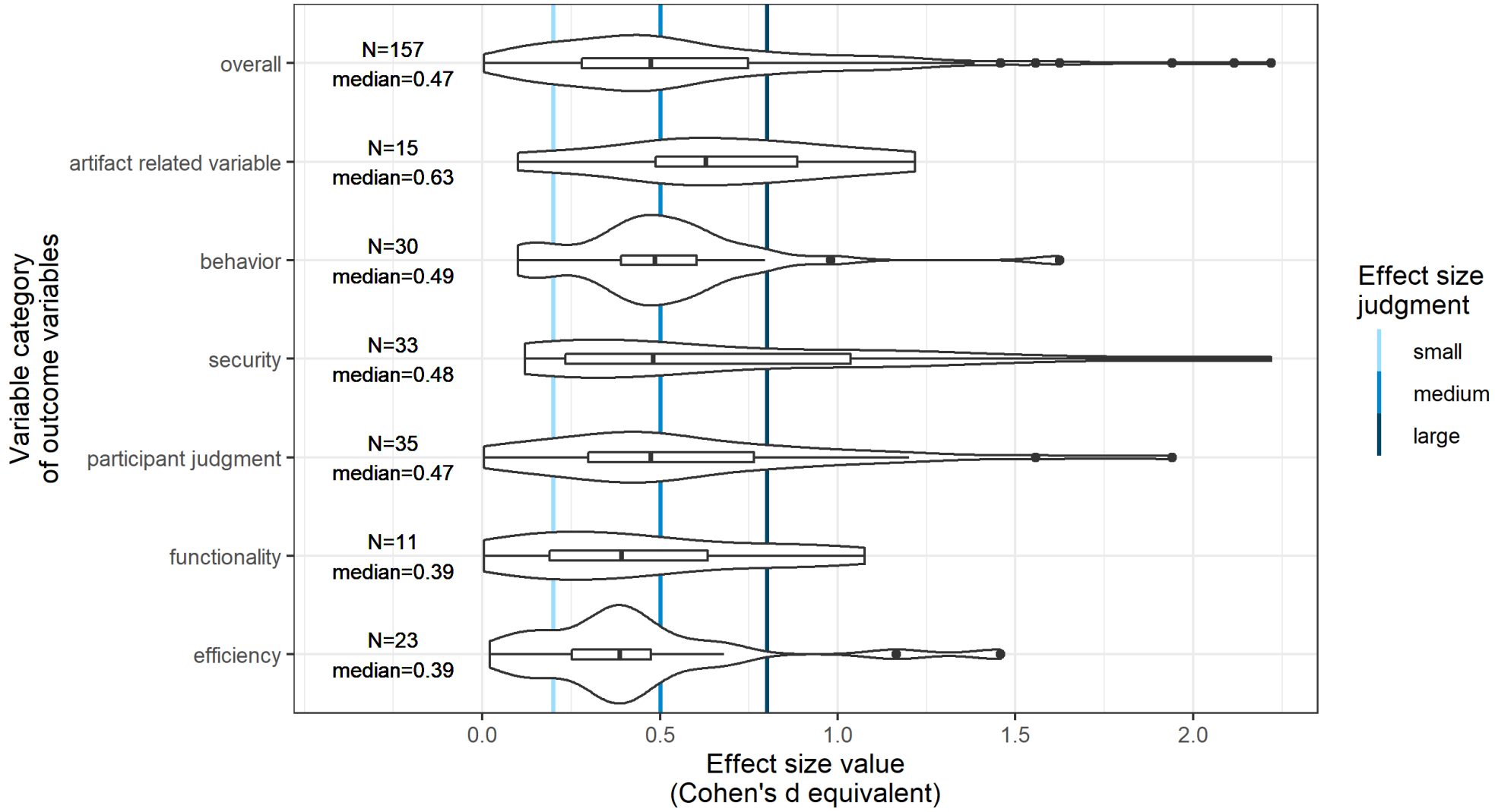


Recommendation

Make fully anonymized data sets available when needed.



Effect sizes in DCUS



Behavioural Security Group

UNIVERSITÄT BONN

Extension: Effect sizes at CHI

Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation

Anna-Marie Ortloff
University of Bonn
Bonn, Germany
ortloff@cs.uni-bonn.de

Florin Martius
University of Bonn
Bonn, Germany
martius@cs.uni-bonn.de

Mischa Meier
Fraunhofer FKIE
Bonn, Germany
mischa@mmisc.de

Theo Raimbault
University of Bonn
Bonn, Germany
s6thraim@uni-bonn.de

Lisa Geierhaas
University of Bonn
Bonn, Germany
geierhaa@cs.uni-bonn.de

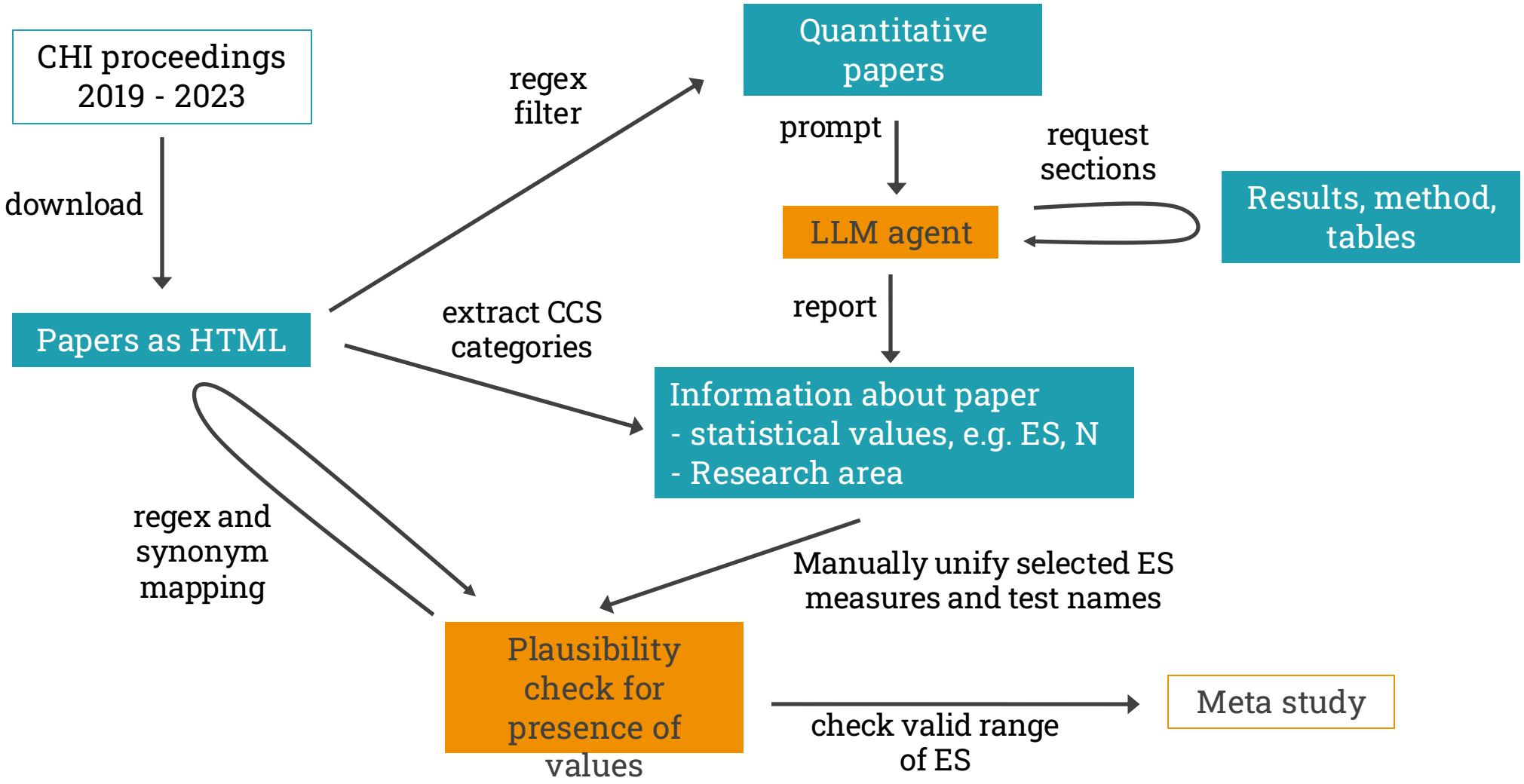
Matthew Smith
University of Bonn
Bonn, Germany
Fraunhofer FKIE
Bonn, Germany
smith@cs.uni-bonn.de

CHI‘25

Community Recommendation

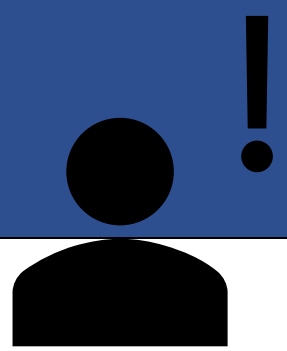
Find a way to standardize and make machine-readable the research output / statistical results.

Extracting Statistical Values from CHI publications

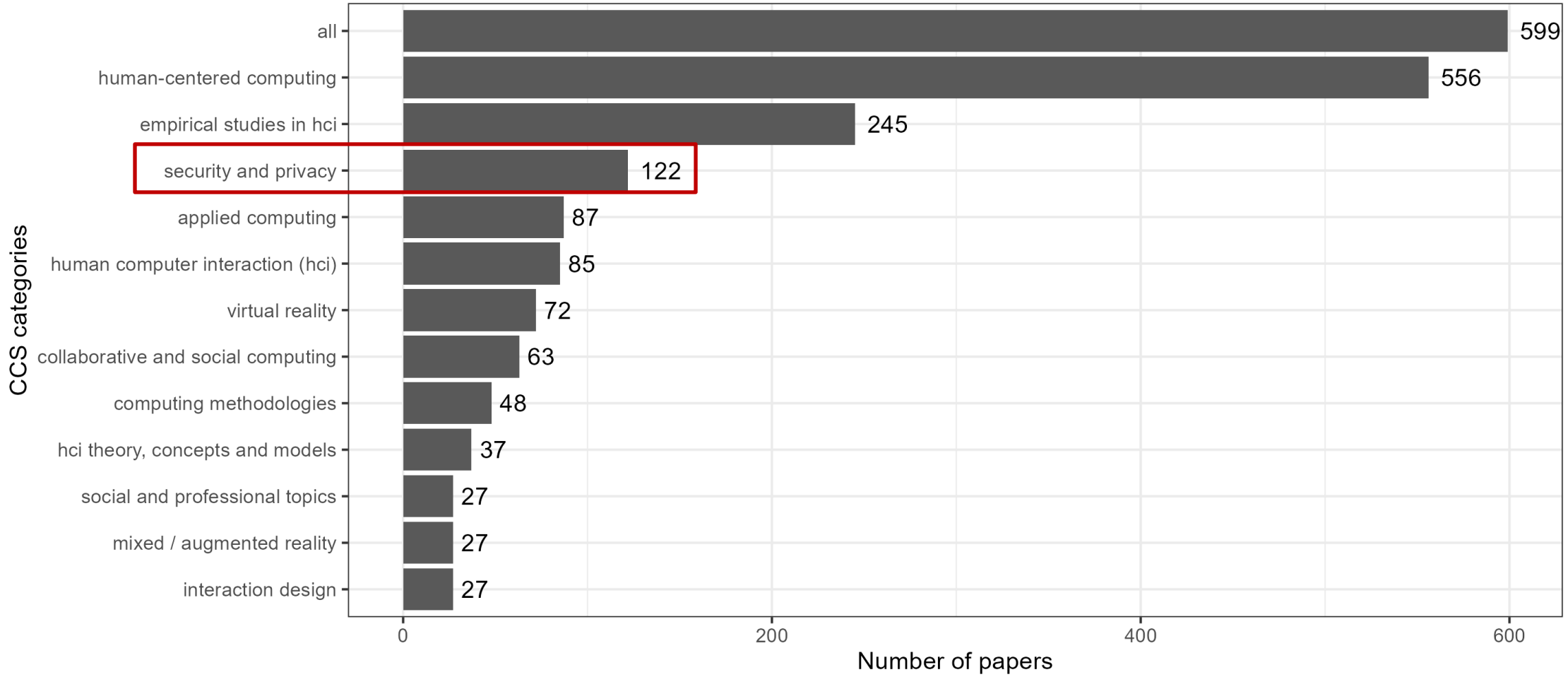


Recommendation

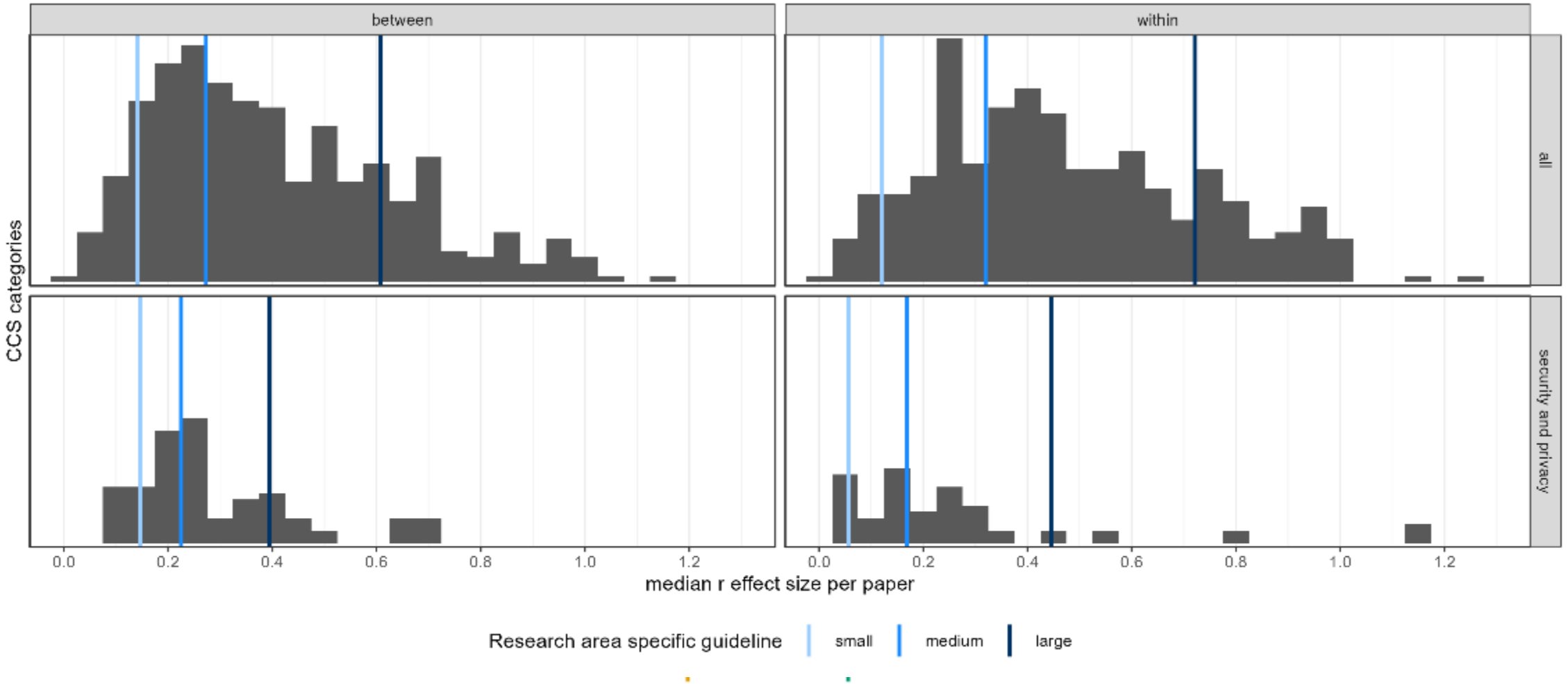
Report all effect sizes, even non-significant ones.



Overview of Quantitative CHI papers 2019 - 2023

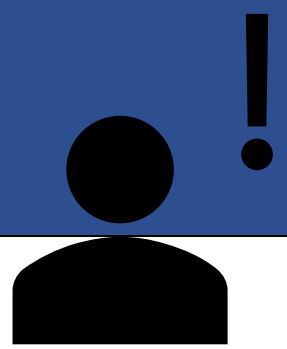


Context-specific effect size guidelines



Recommendation

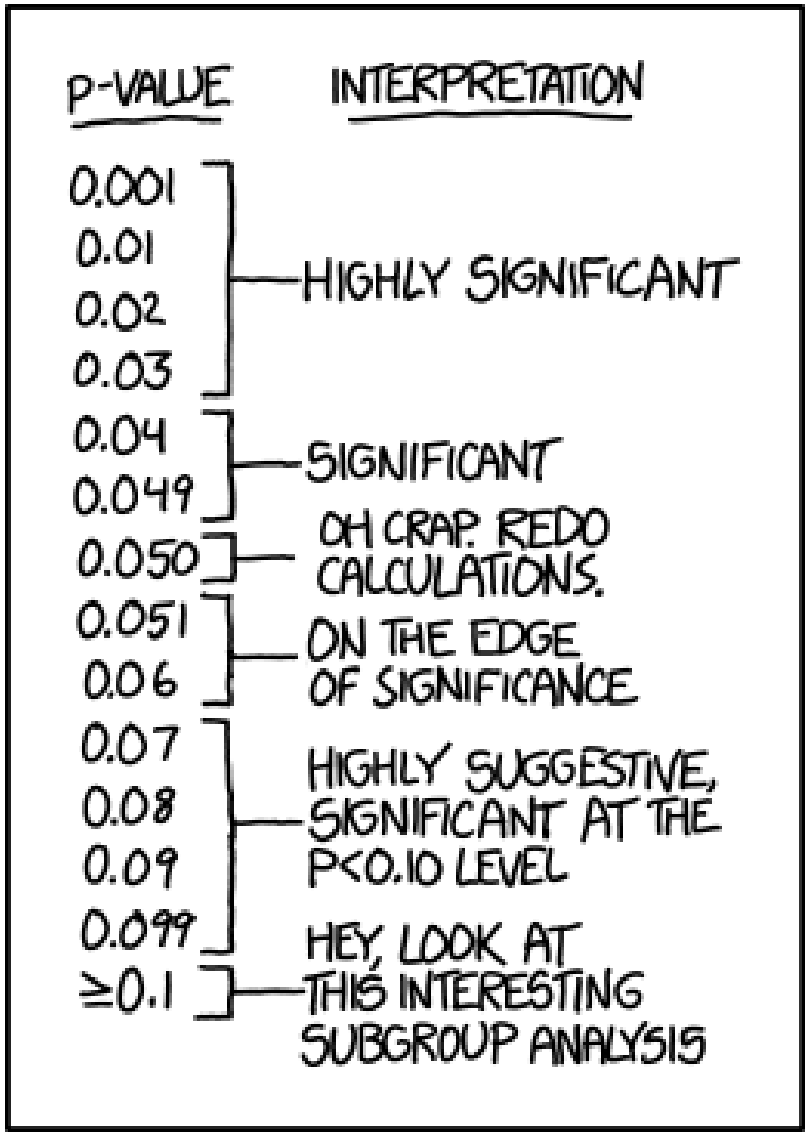
Take into account context
when interpreting effect sizes.



Behavioural Security Group

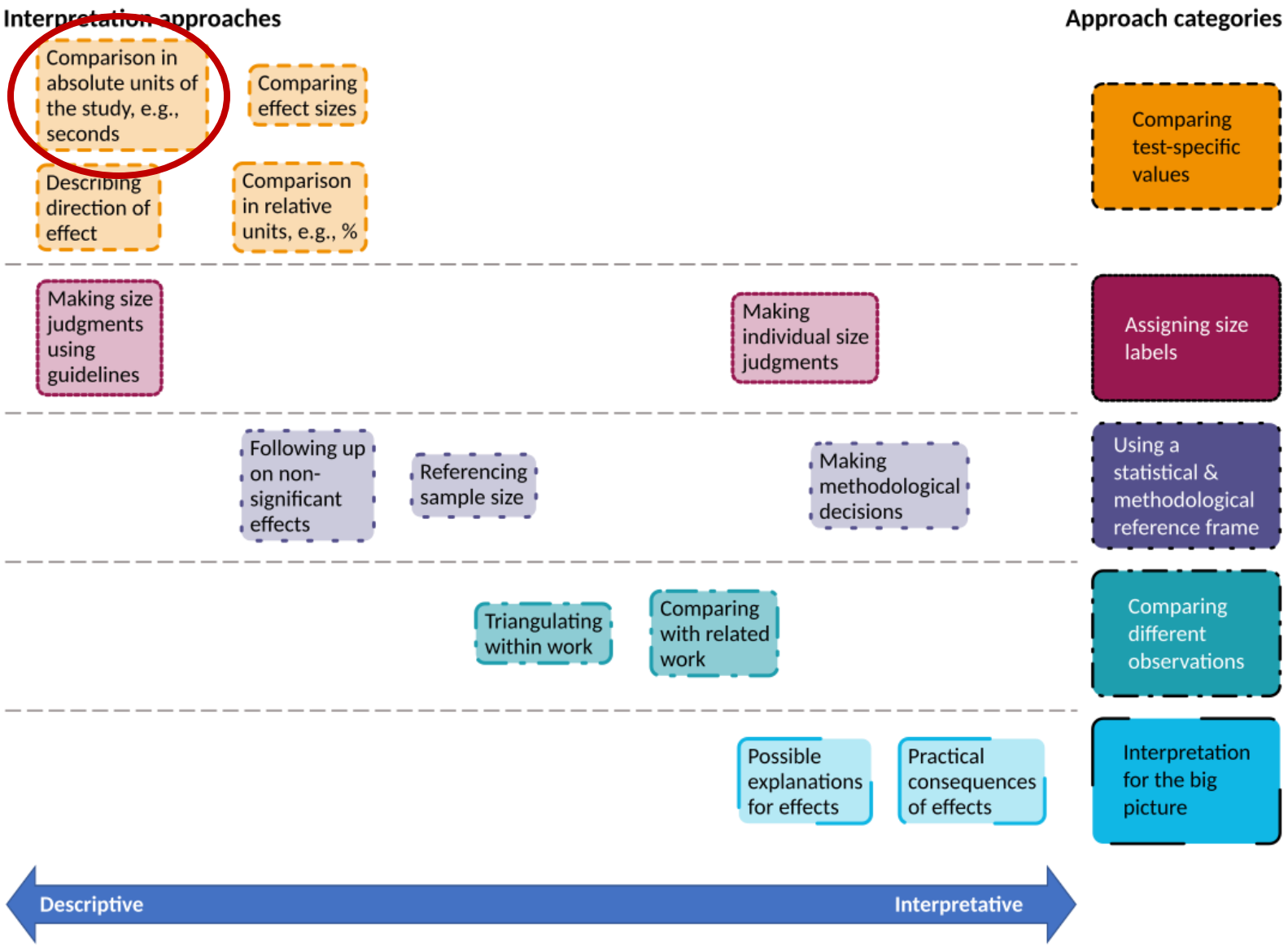
UNIVERSITÄT BONN

What about interpretation of effects?



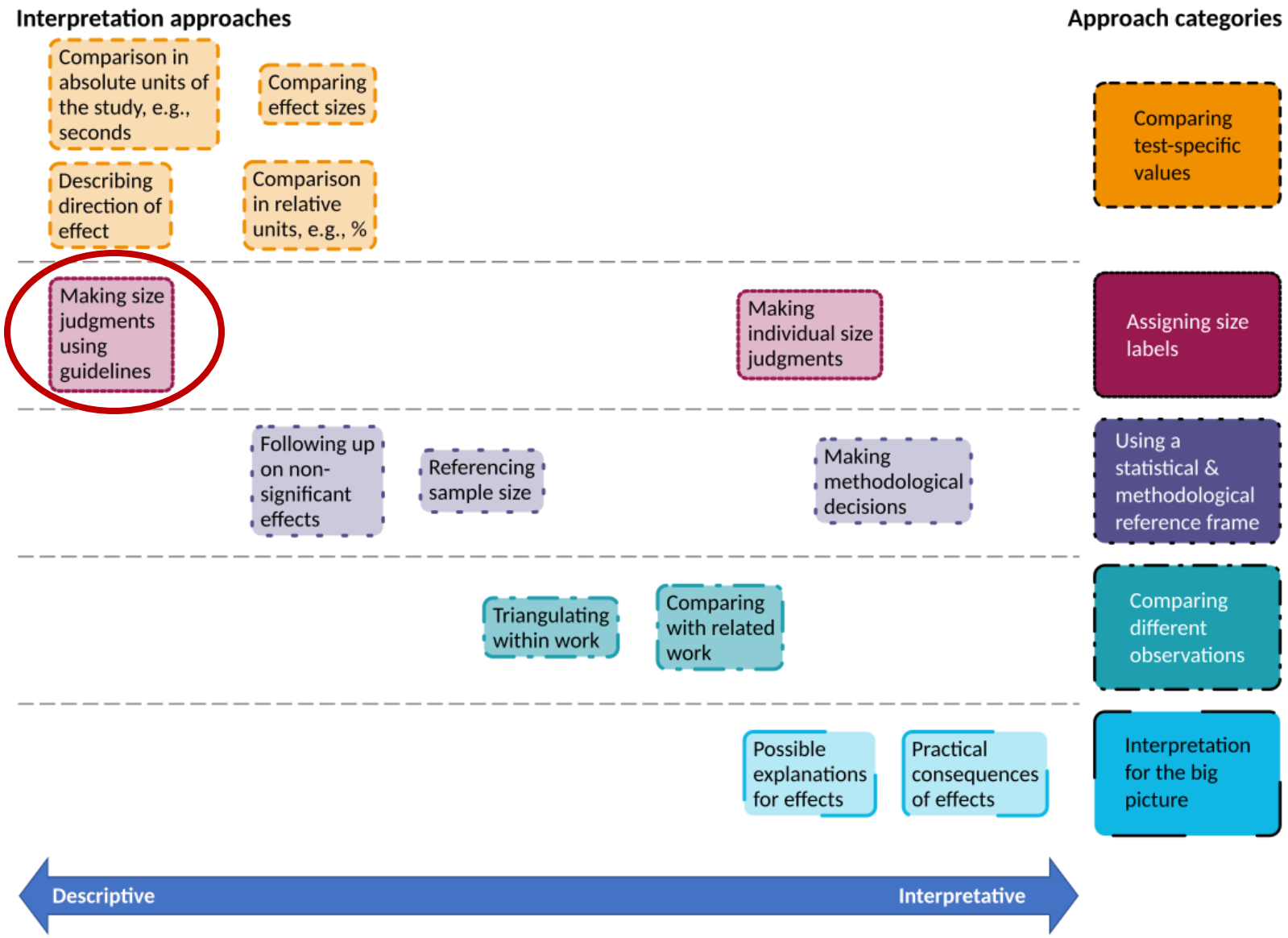
https://www.xkcd.com/1478/

Source: Publications



"[...] significantly increased the number of words in a sentence after which suggestions were requested – **by about 1.5 words**"
– Dang et al.: Choice over Control (https://dl.acm.org/doi/full/10.1145/3544548.3580969)

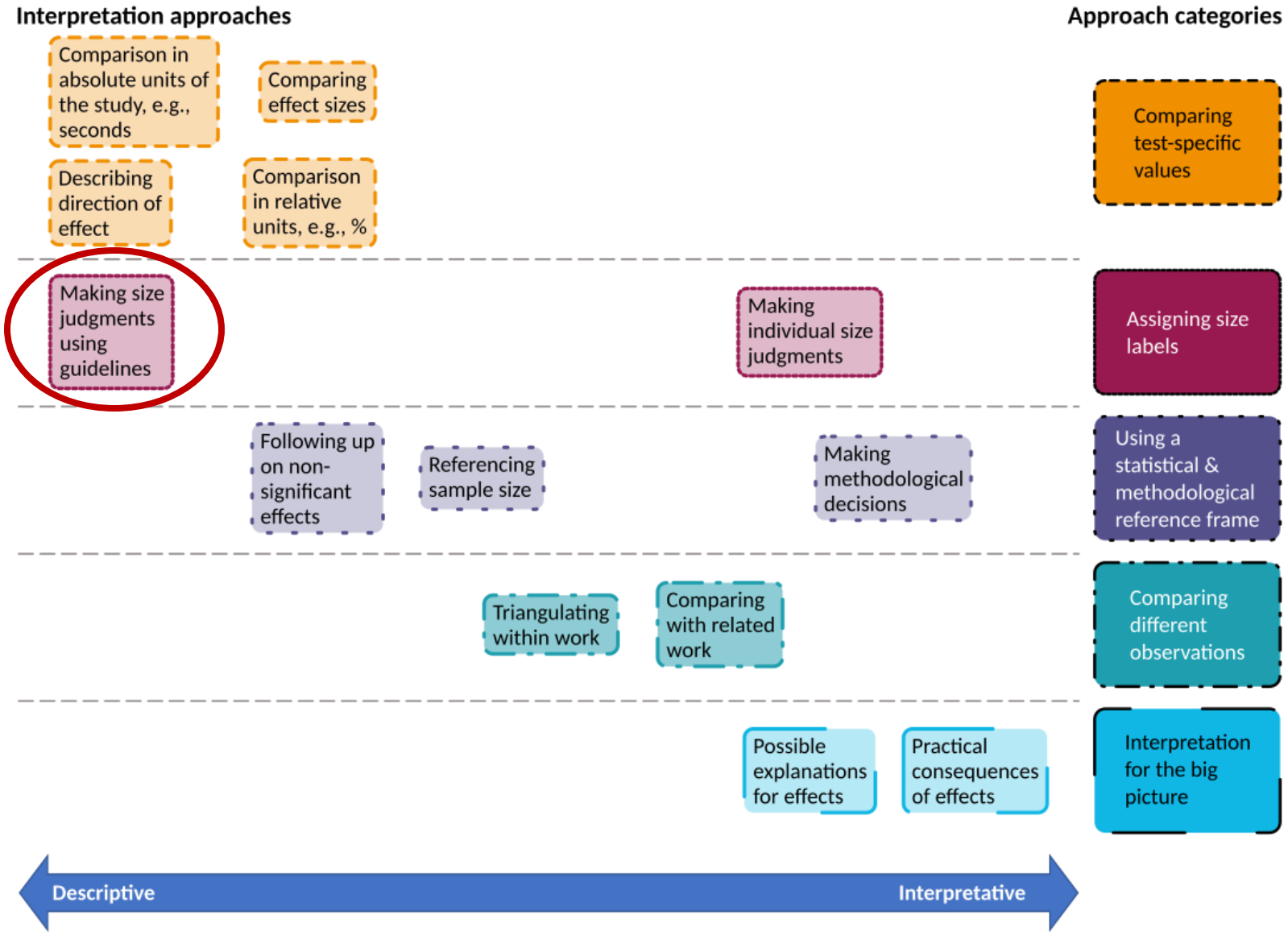
Source: Publications

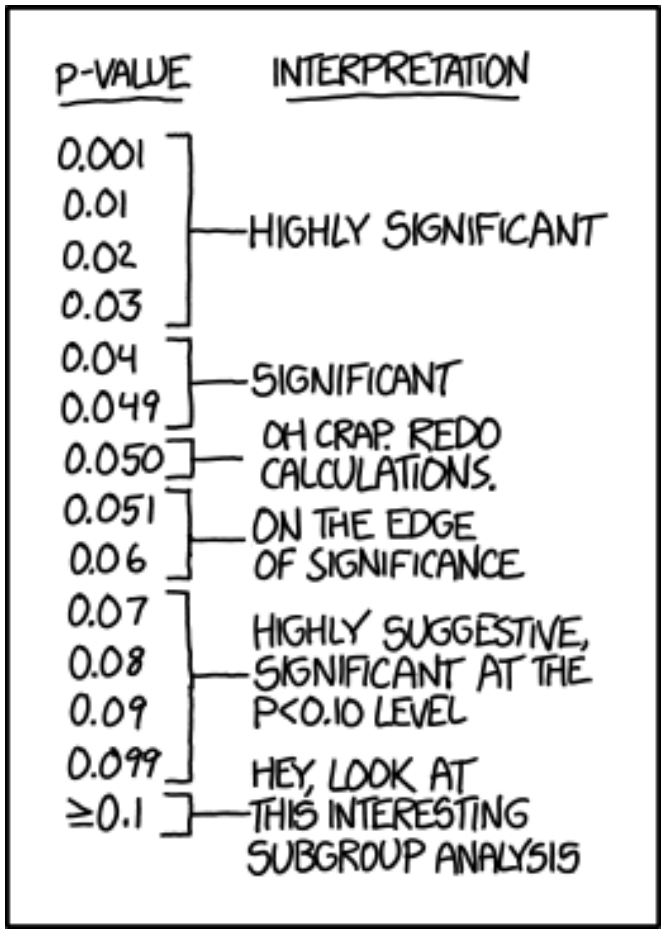


„Both plank (static core) and press-ups (dynamic upper) showed only non-significant, small effect
sizes (Cohen's d = 0.239 and 0.298).”

- Clarke et al.: FakeForward (https://dl.acm.org/doi/abs/10.1145/3544548.3581100)

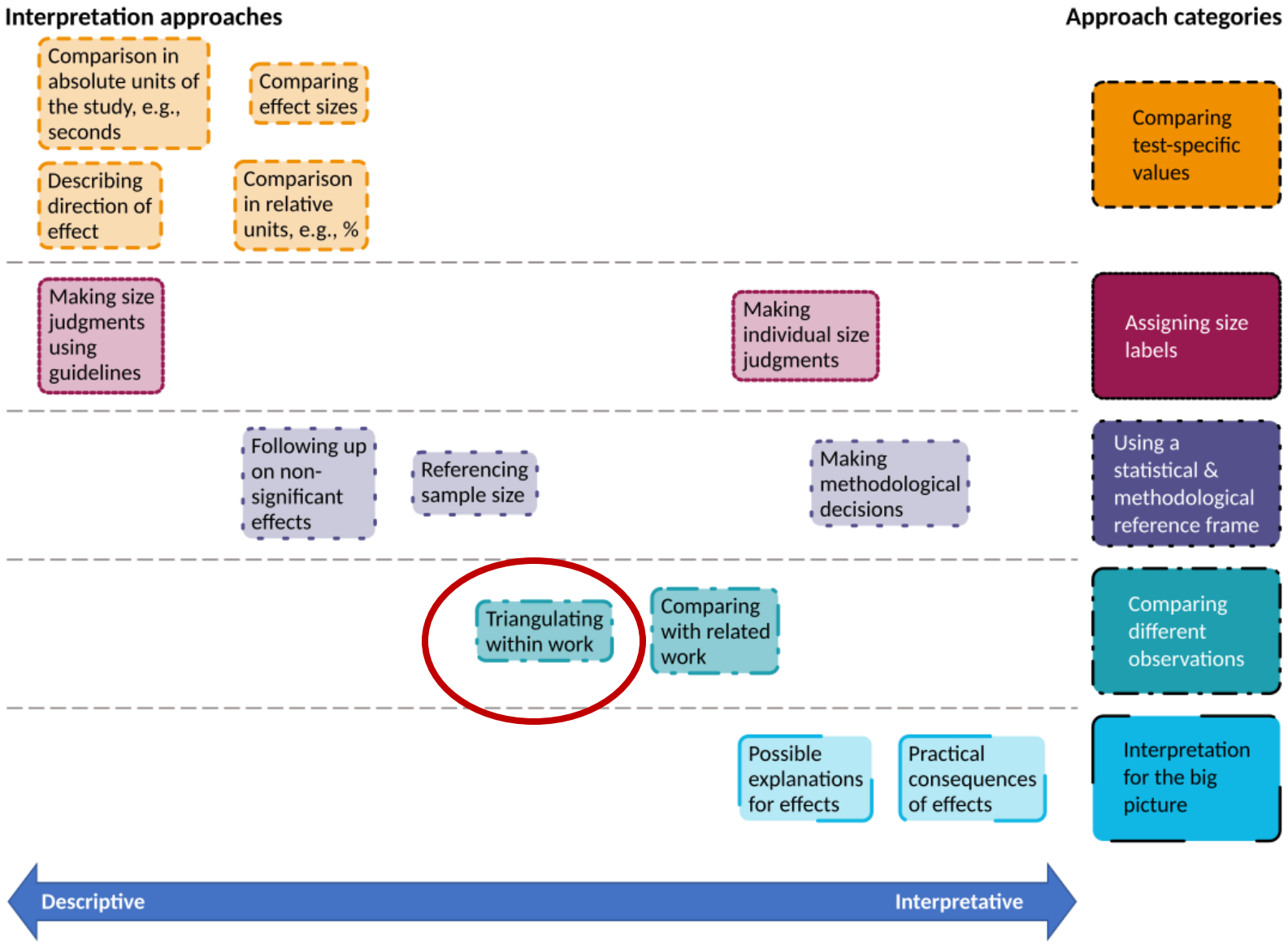
Source: Publications





https://www.xkcd.com/1478/

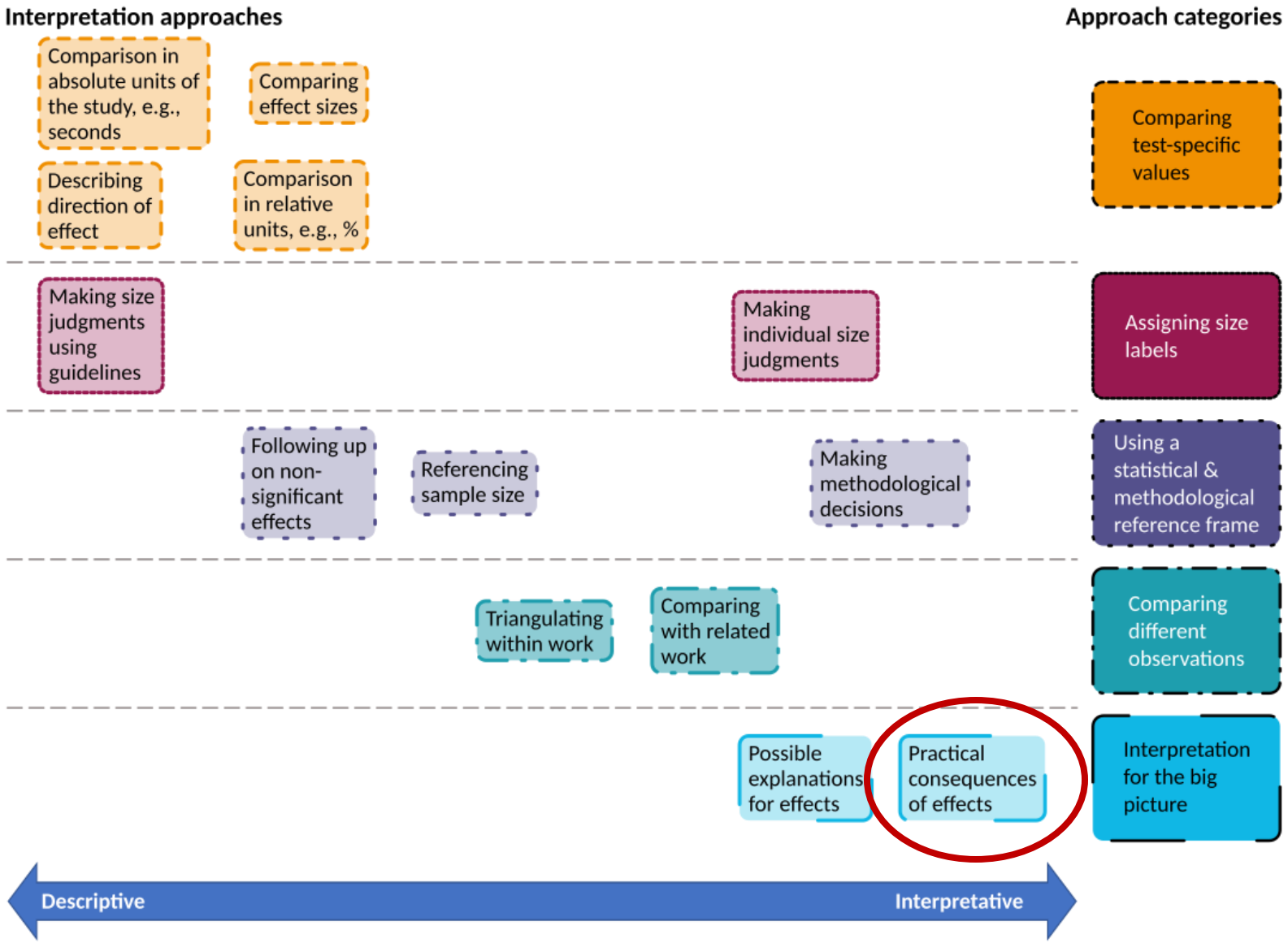
Source: Publications



„We could not draw a definite conclusion about which algorithm was more accurate because GlanceWriter had a lower error rate but it also had a higher number of error correction actions."

- Cui et al.: GlanceWriter (https://dl.acm.org/doi/abs/10.1145/3544548.3581100)

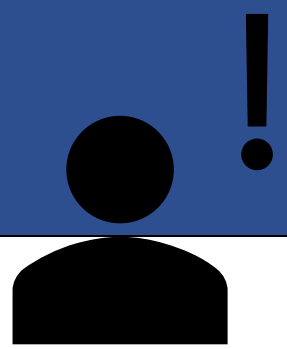
Source: Publications



"While eye-tracking seems to be the favourite visual attention cue, not all VR headsets have eye-tracking capabilities, and perhaps future cheap VR headsets will never incorporate such capabilities. Our results show that […], **cheap VR headsets using bi-directional CoV can still lead to the same amount of joint attention**, therefore, being effective."
- Bovo et al.: Speech-Augmented Cone-of Vision (https://dl.acm.org/doi/full/10.1145/3544548.3581283)

Recommendation

Interpret and Discuss Effect Sizes.



Source: Interviews and Surveys

A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes

Anna-Marie Ortloff
University of Bonn
Bonn, Germany
ortloff@cs.uni-bonn.de

Julia Angelika Grohs
University of Bonn
Bonn, Germany
s6jugroh@uni-bonn.de

Simon Lenau
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
lenau@cispa.de

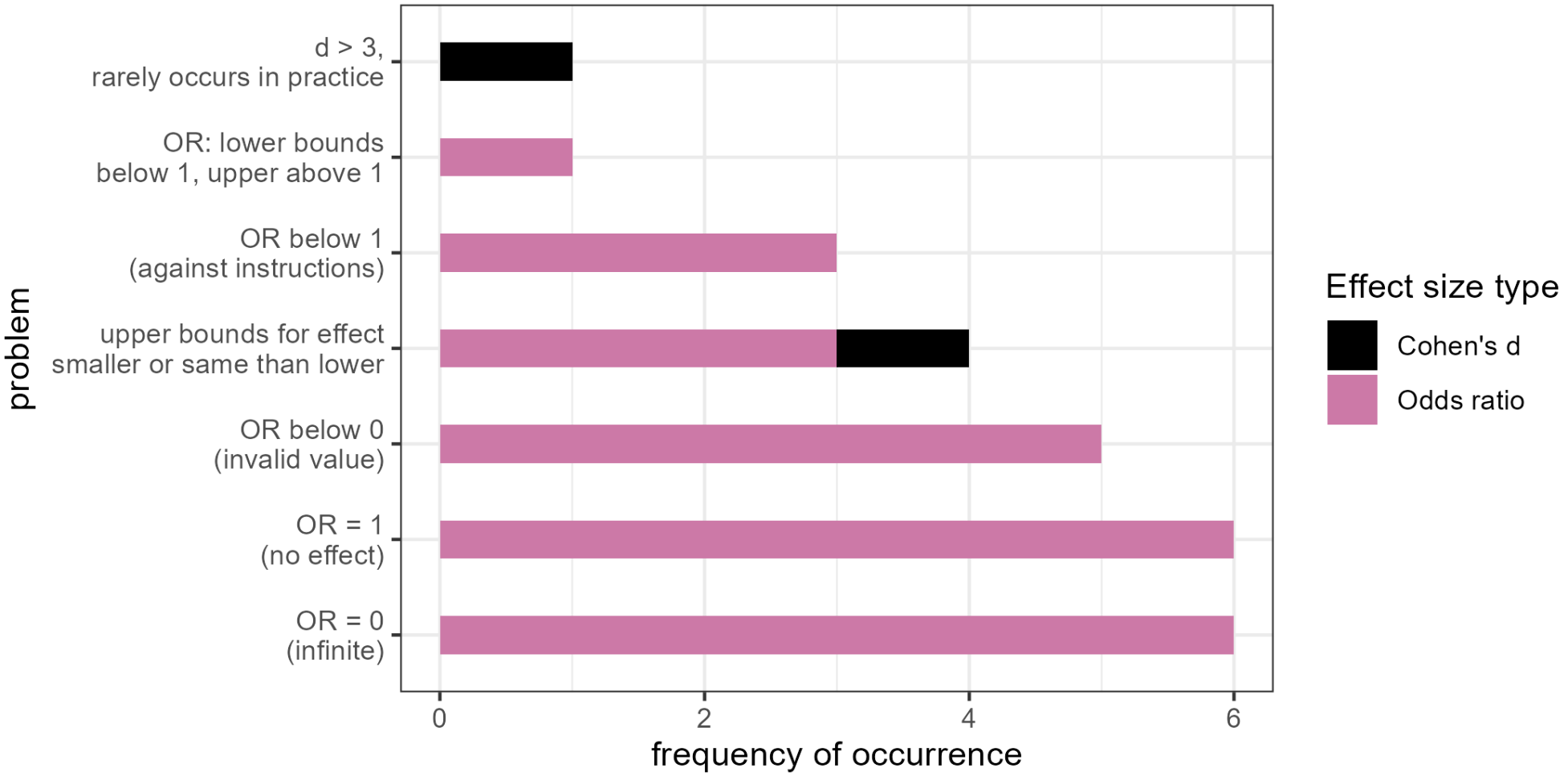
Matthew Smith
University of Bonn
Bonn, Germany
Fraunhofer FKIE
Bonn, Germany
smith@cs.uni-bonn.de

CHI'25

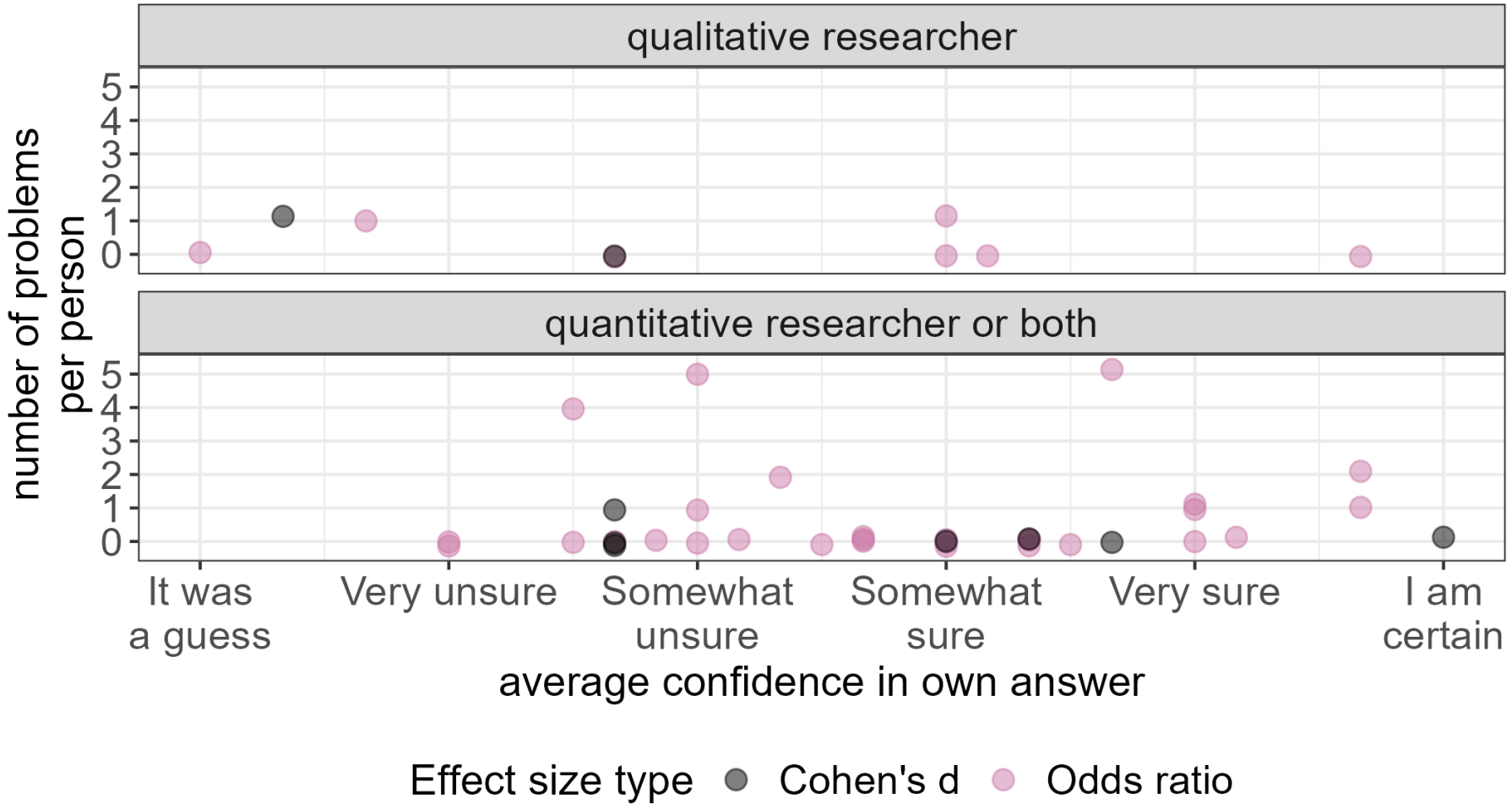
Vignette – Password Manager Example

A study investigated the difference between the adoption of **a password manager**, in a baseline group that received a general introduction to **the password manager** and an intervention group that additionally was informed that **using a password manager was the top recommendation made by security experts.** 2100 Participants used the **password manager** during the study. Two weeks after the end of the study, they were asked whether they were still using the **password manager** or not.	RQ and study design
511 out of 1050 (**48.7%**) of those who received the intervention were still using it. **224** out of 1050 (**21.3%**) of those who did not see the intervention were still using it	Descriptive Stats
Fisher's exact test showed that this difference was statistically significant ($p<0.001$, odds ratio = **3.50,** 95% CI=[**2.88, 4.25**]). The effect size (Odds ratio) is **3.50**.	Inferential Stats
This means that the odds for the participants to continue to use the formatting tool in the intervention group were **3.50** times higher than for participants to continue to use the **password manager** in the baseline group.	Effect size explanation

Misconceptions about OR and Cohen‘s d

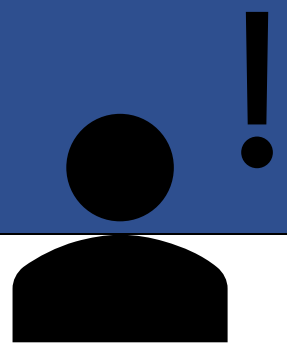


Understanding of Effect Size Measures



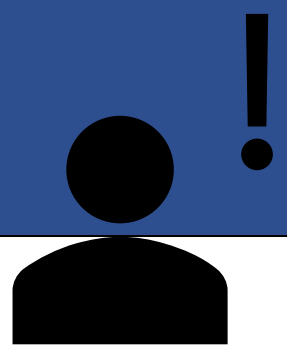
Recommendation

Explain effect size measures.



Recommendation

Report standardized **and** non-standardized effect sizes.



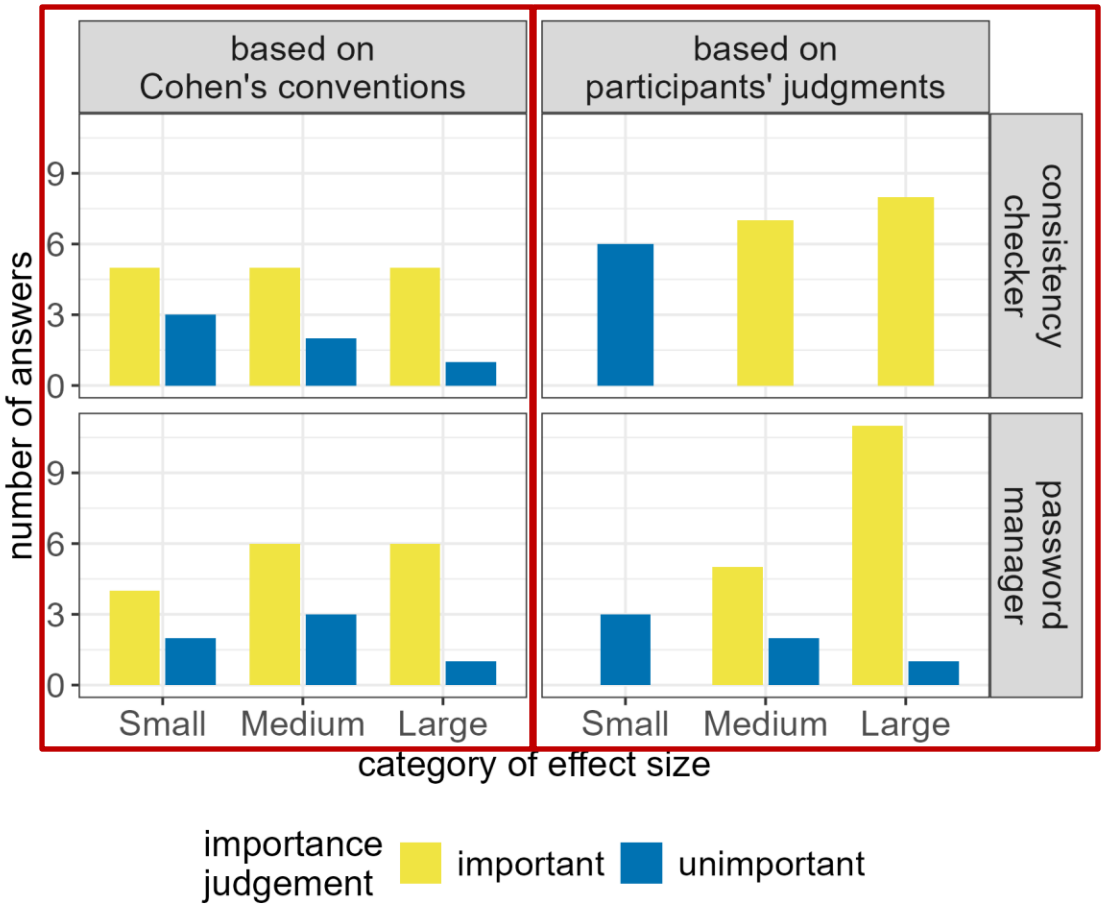
Influencing Factors on Interpretation of Effect Size

- Size
- Context
- Point of view
- Other numerical values in the vignette

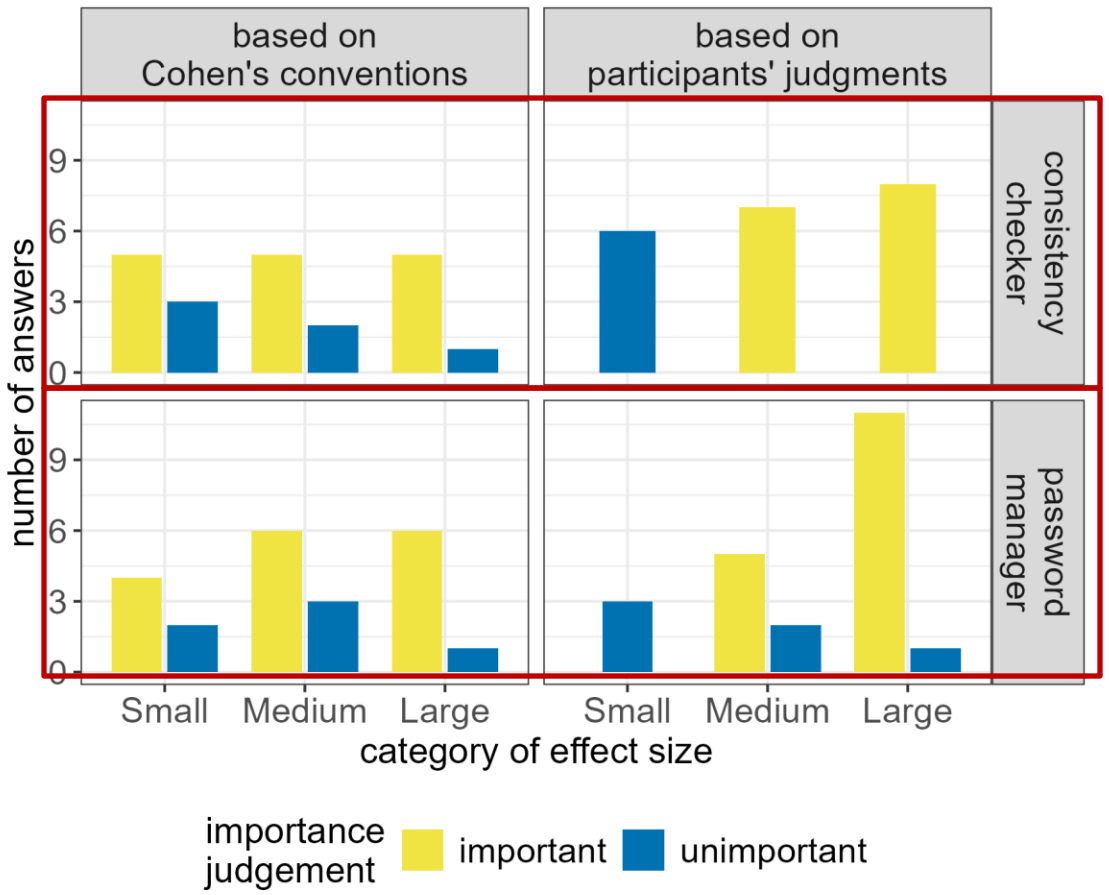
Influencing Factors on Interpretation of Effect Size

- **Size**
- **Context**
- Point of view
- Other numerical values in the vignette

Judgment Based on Size



Judgment Based on Context



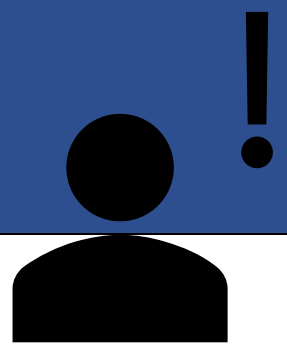
"[**Importance**] depends on the topic and the population. Using a password manager is not the same as curing cancer." (from survey)

On effect **size**:

"Comparing an elephant to the earth, it's small, but compared to a mouse it is big" (from interview)

Recommendation.

Interpret and Discuss Effect Sizes.



Recommendations for Reporting

Reporting Statistical Results

- Specify if you are using confirmatory or exploratory analysis methods
- Report all effect sizes - even non-significant ones
- Report standardized and non-standardized ES
- Interpret and discuss ES
- Explain ES measures

Make Reporting Usable for Future Work

- Report complete and appropriate descriptive statistics
- Make fully anonymized data sets available when needed
- Find a way to standardize statistical research output and make it machine-readable

Recommendations up for Discussion

- Move test statistics to supplementary material

- Consider not reporting p-values and focus on confidence intervals instead

Community Recommendations

Develop / Use Reporting Guidelines.