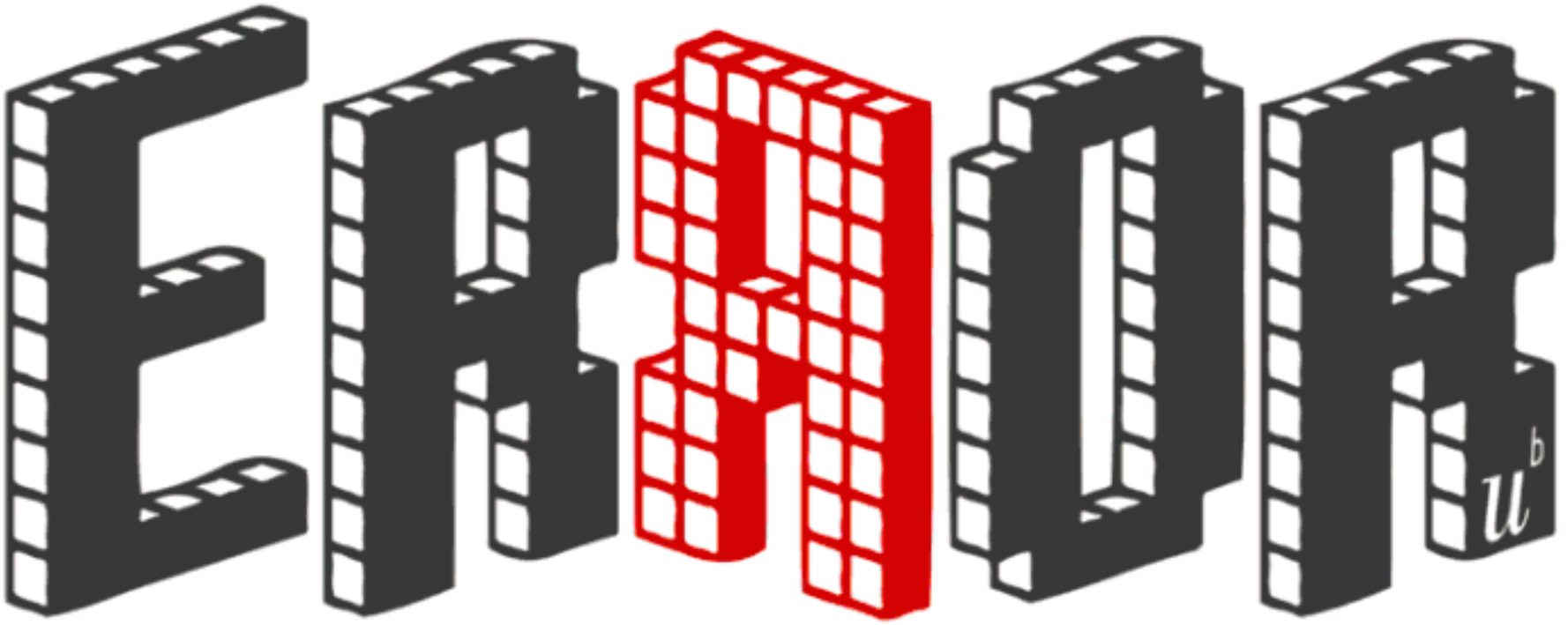


ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

UNIVERSITÄT BERN

Malte Elson

University of Bern

Ruben Arslan

University of Leipzig

Ian Hussey

University of Bern

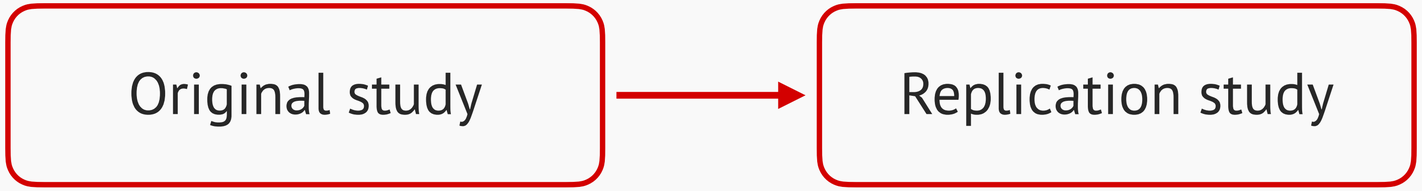
Jamie Cummins

University of Bern

“Like technology companies, stakeholders in science must realize that making error detection and correction part of the scientific landscape is a sound investment.”

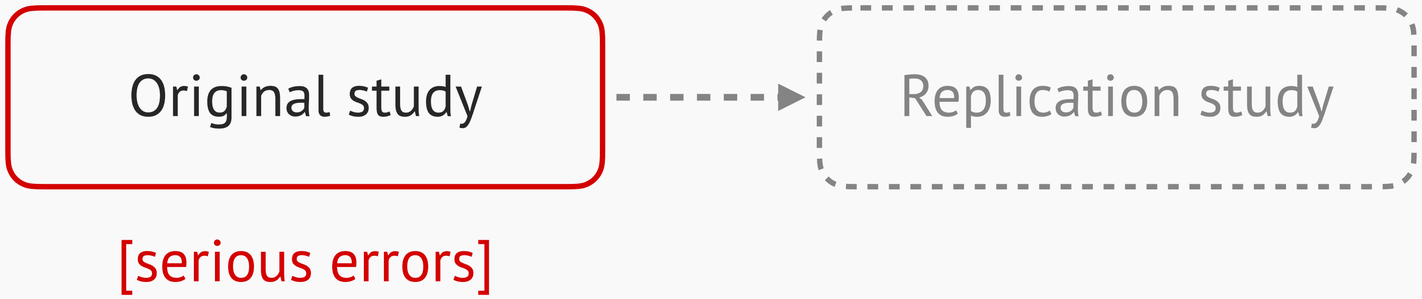
Elson & Stapel (2024) *Nature*

Replications are expensive



Replications are expensive

Error inspections can be a useful prior step



We ask too much of peer-review

✗ One-shot, little redundancy
Despite imperfect reliability & evolving knowledge

✗ Generally non-technical

✗ Surprising little error checking
for an industry built on discovering truth

Make it easy
Develop new methods and tools to make
error detection and trustworthiness assessment easier

Make it normative
Master’s degree course in error detection

Make it rewarding
ERЯOR: A bug bounty program for science

UNIVERSITÄT BERN

Make error detection easier

xxxxxxxxxxxxxxxxxxxxxxxxxx

01.

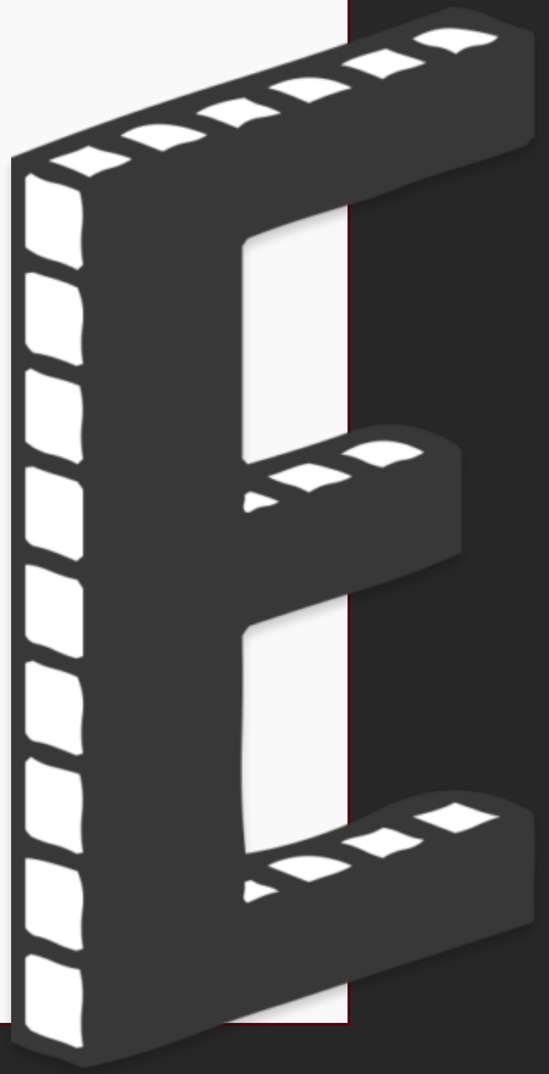
01.

Error / trustworthiness checks

Expertise problem

Beck Depression Inventory (BDI-II)

$N = 23, M = 20.70, SD = 3.40$



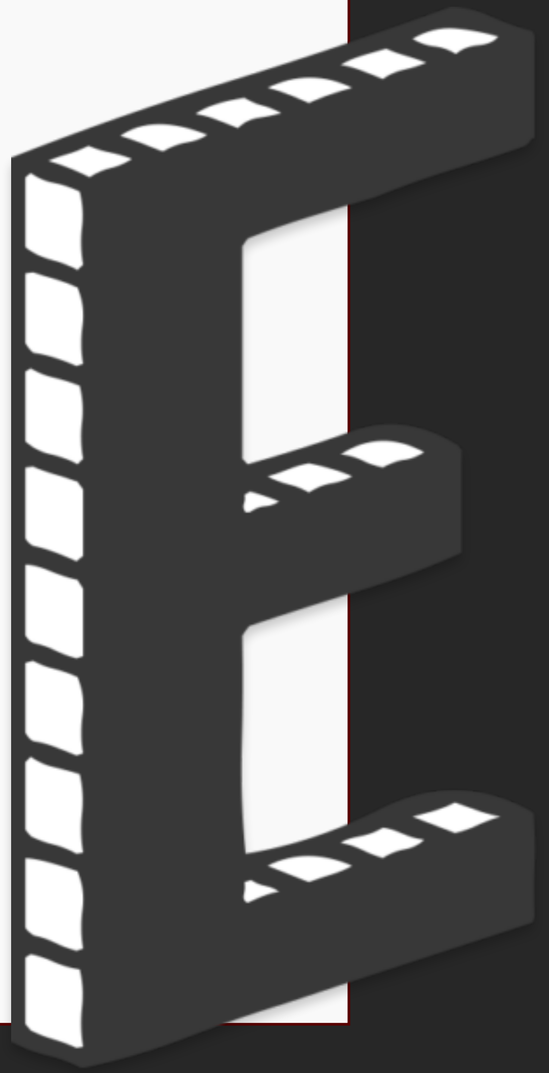
01.

Error / trustworthiness checks

Information that is repeated or overlooked

Granularity tests

N = 2, Mean age = 25.3



GRIM & GRIMMER (Brown & Heathers, 2016; Anaya, 2016; Allard, 2018)

01.

Error / trustworthiness checks

Information that is repeated or overlooked

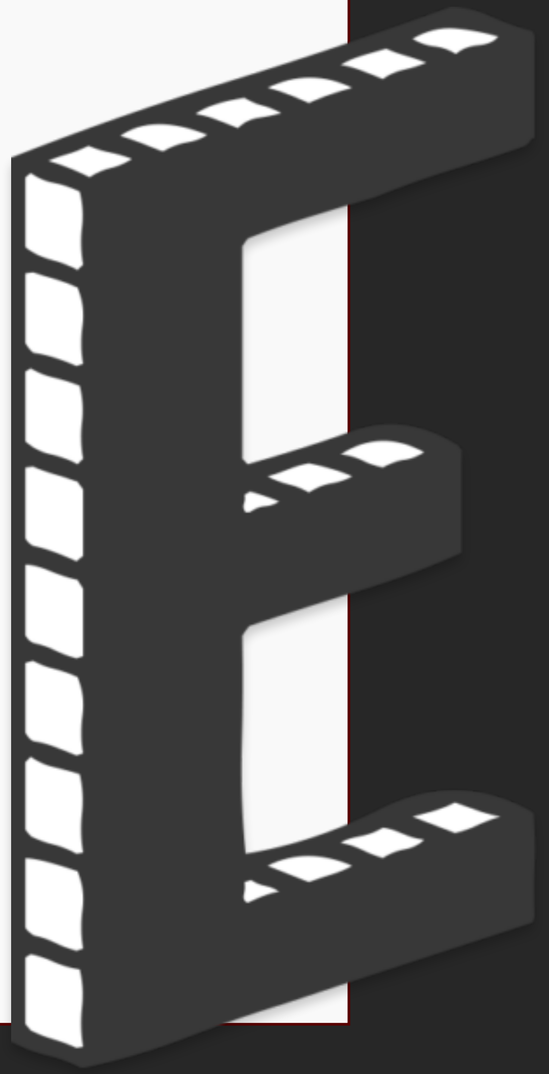
Granularity tests

N = 2, Mean age = 25.3

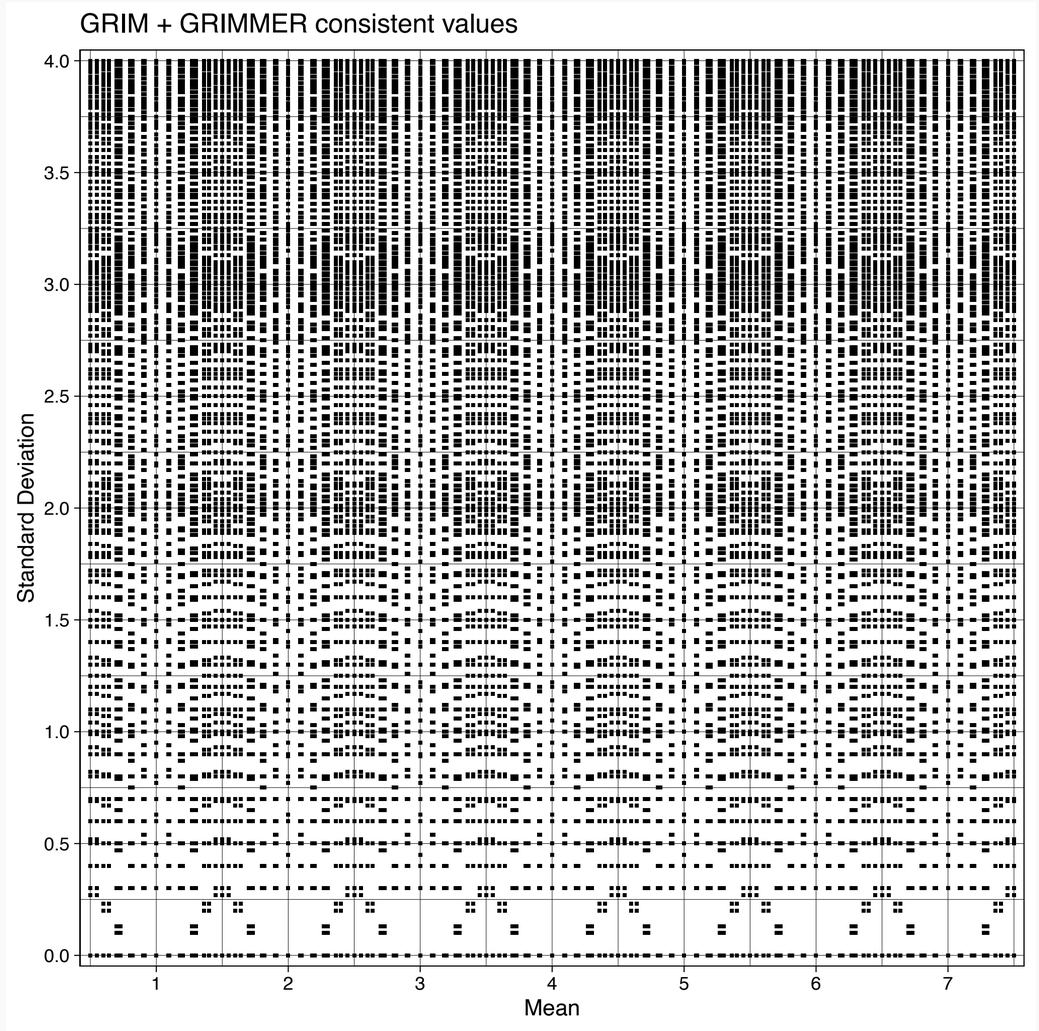
Bounds tests

1-7 Likert scale, *M* = 8.11

Hinted at in SPRITE (Heathers et al., 2018; Wallrich, 2021)

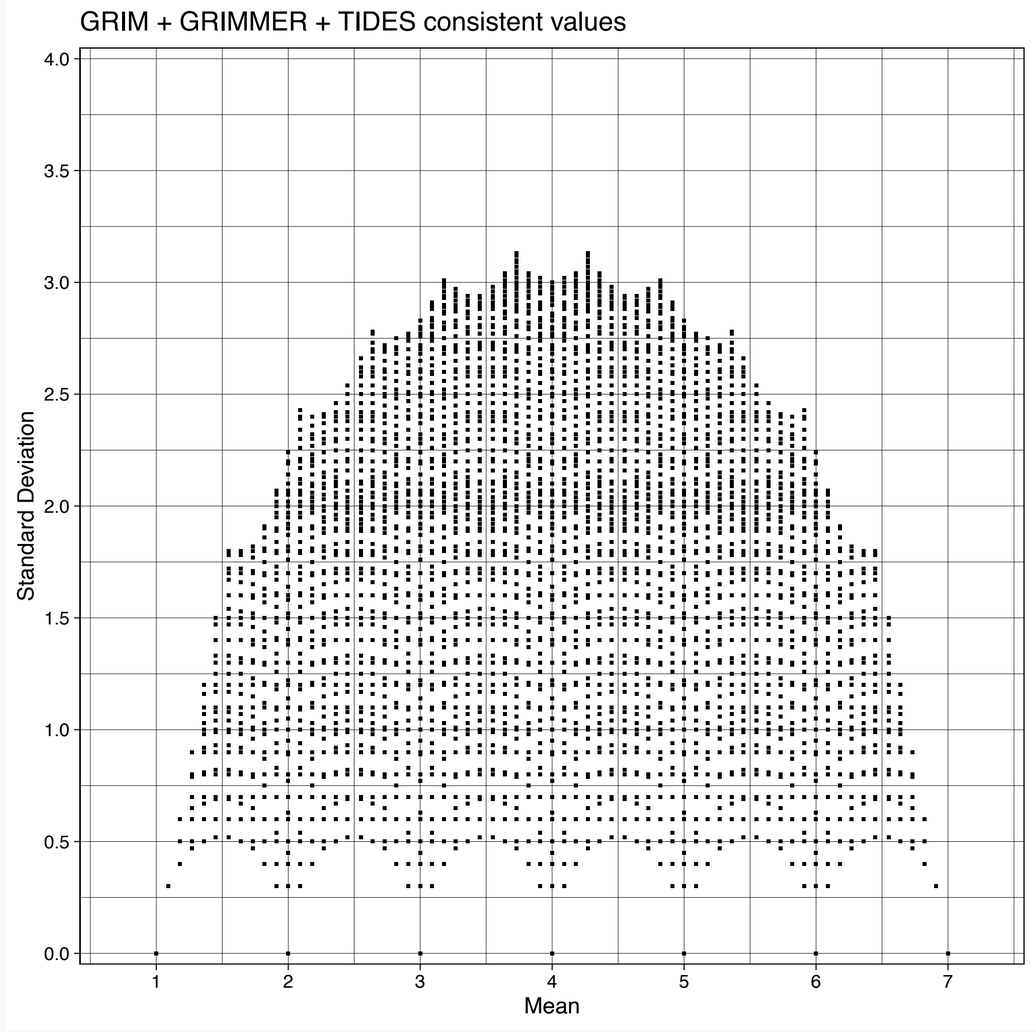


Granularity tests



N=11

Granularity + bounds tests



N=11

‘Umbrella’ plot

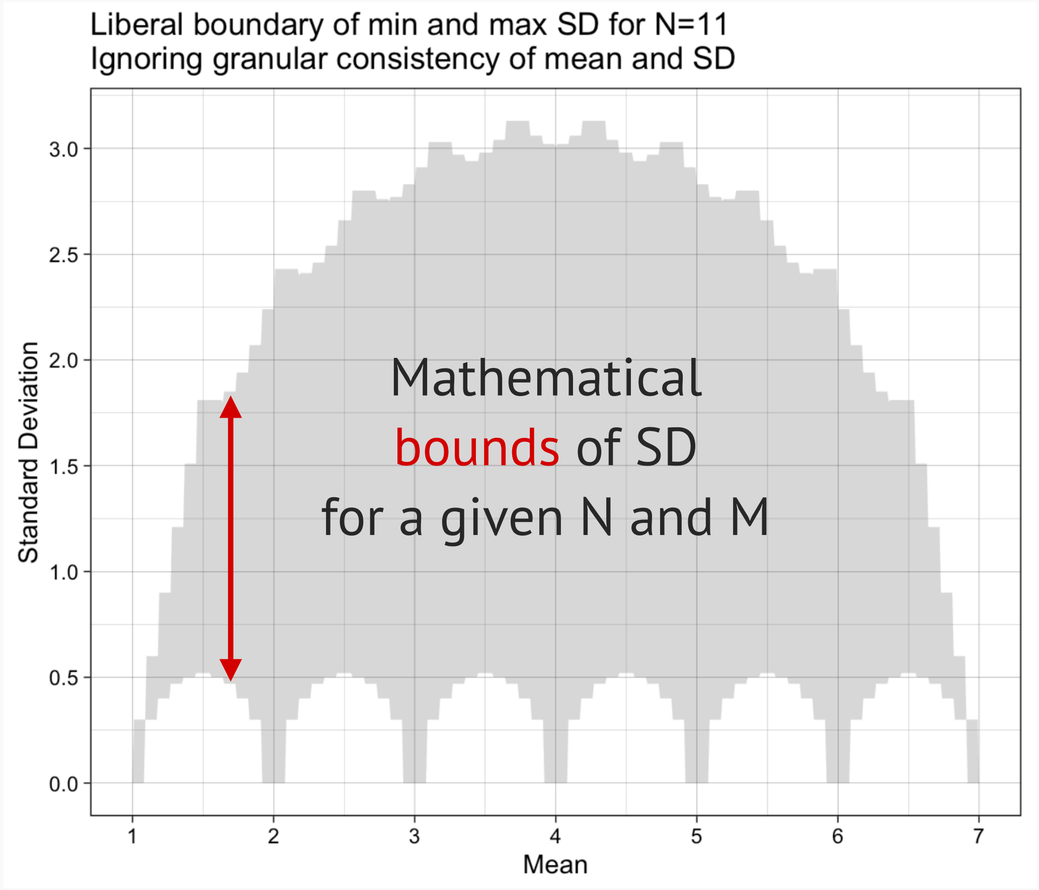
Heathers et al. 2017

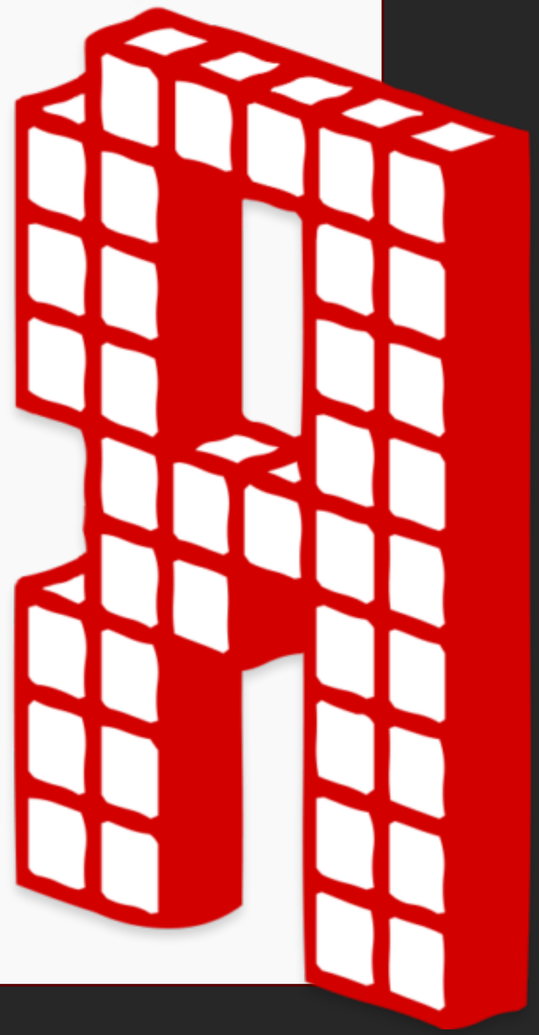
01.

TIDES

Truncation Induced DEpendencies among Summary statistics

Pacman plot





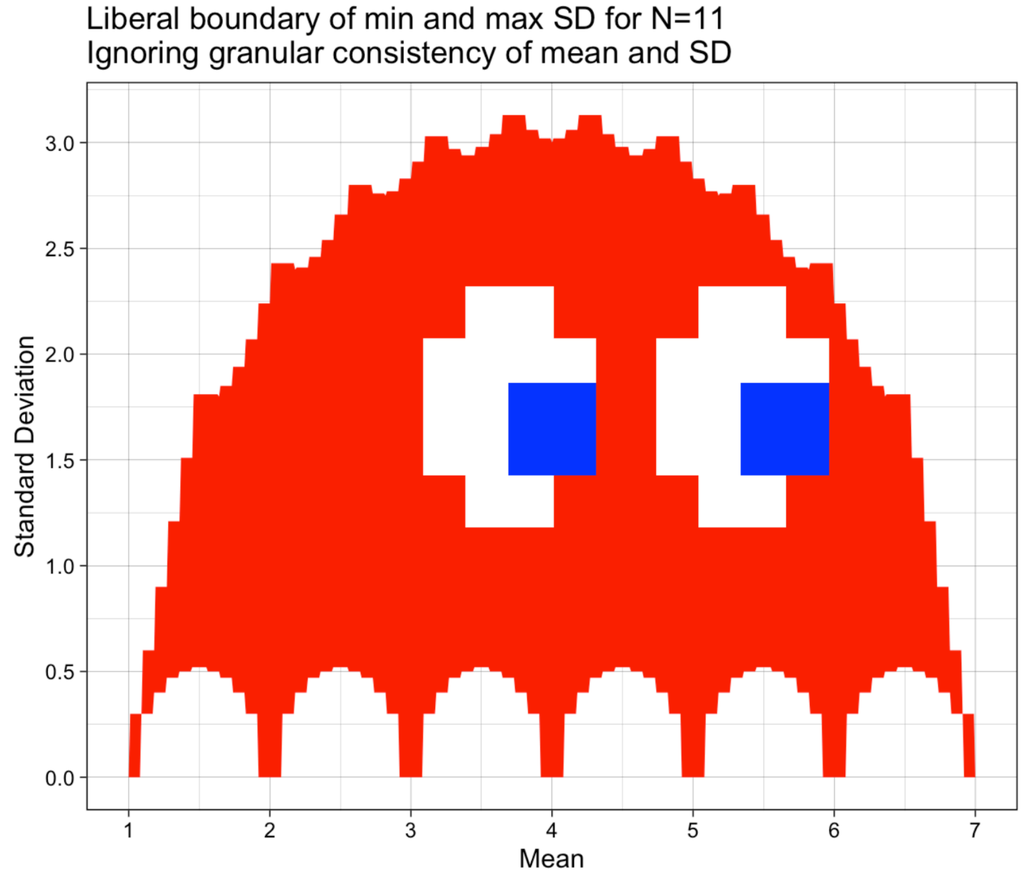
Hussey et al. (in prep)

01.

TIDES

Truncation Induced DEpendencies among Summary statistics

Pacman plot



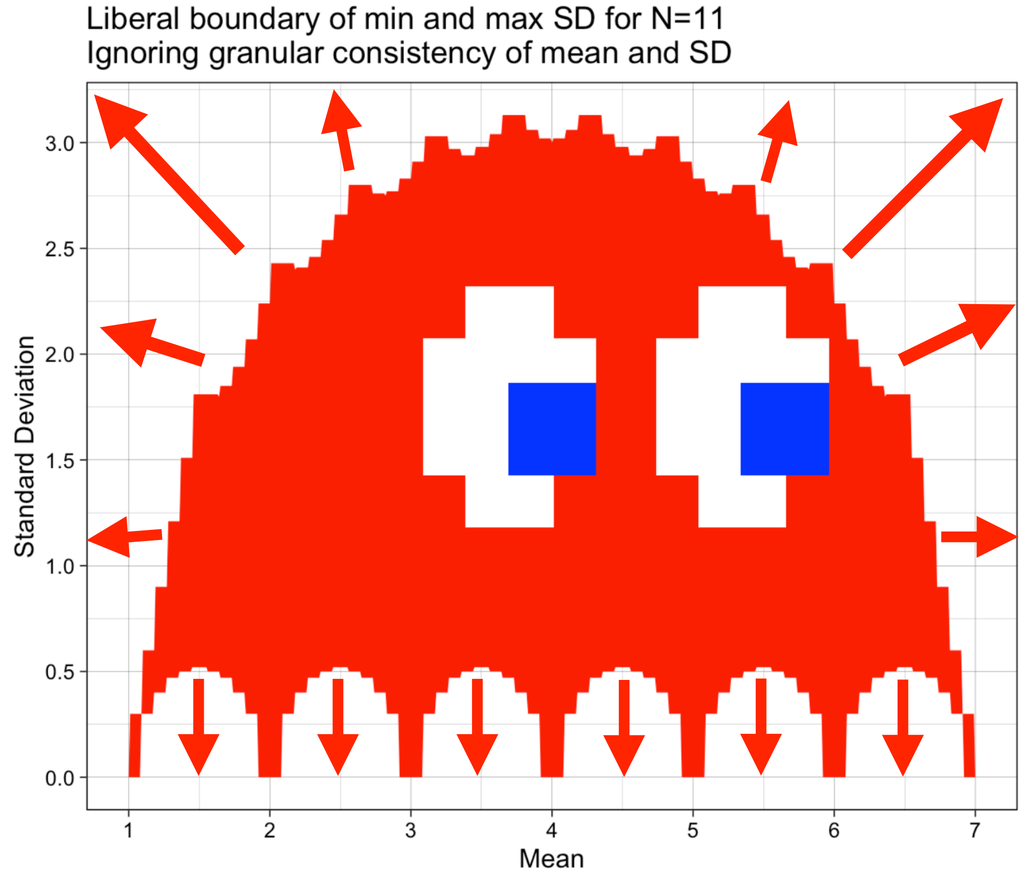
Hussey et al. (in prep)

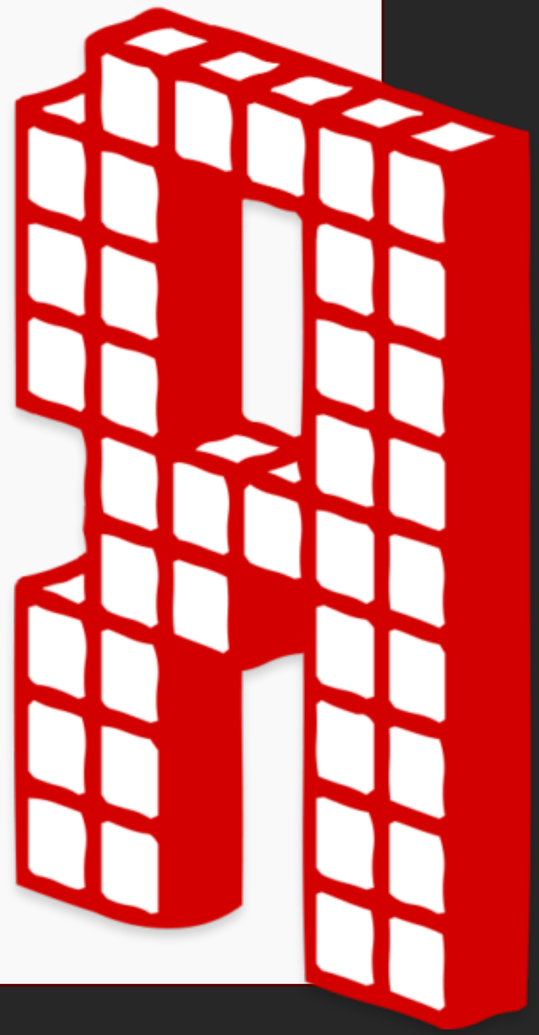
01.

TIDES

Truncation Induced DEpendencies among Summary statistics

Pacman plot





Hussey et al. (in prep)

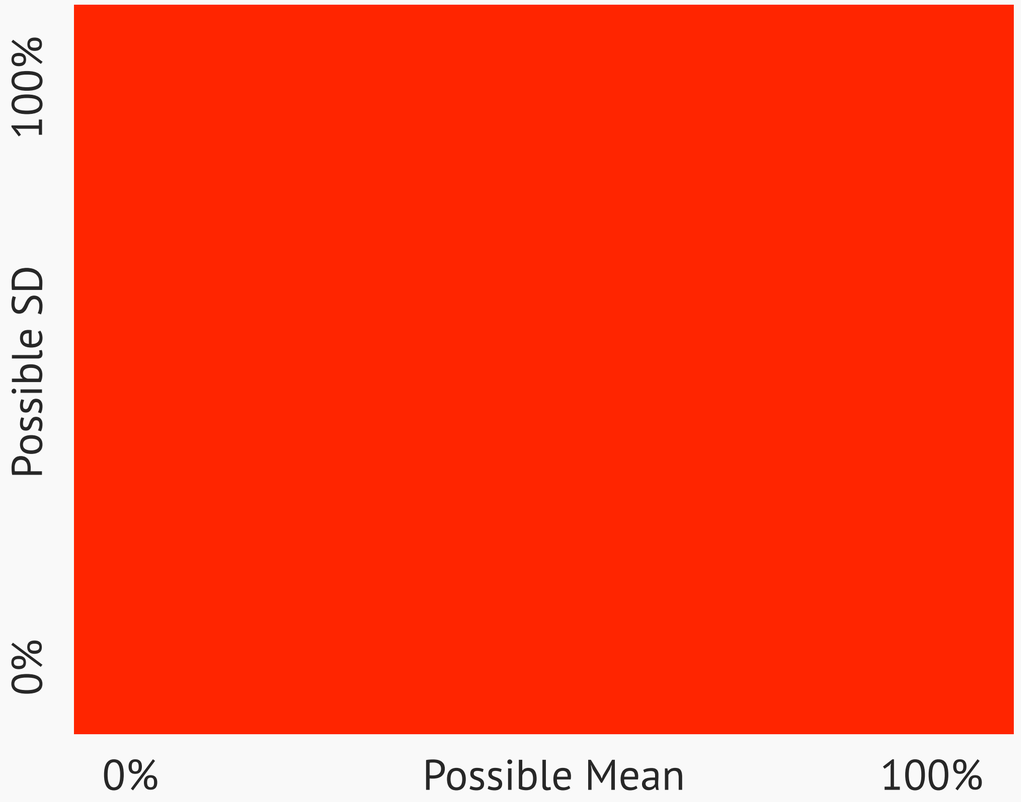
Standardized via Percent-Of-Maximum-Possible

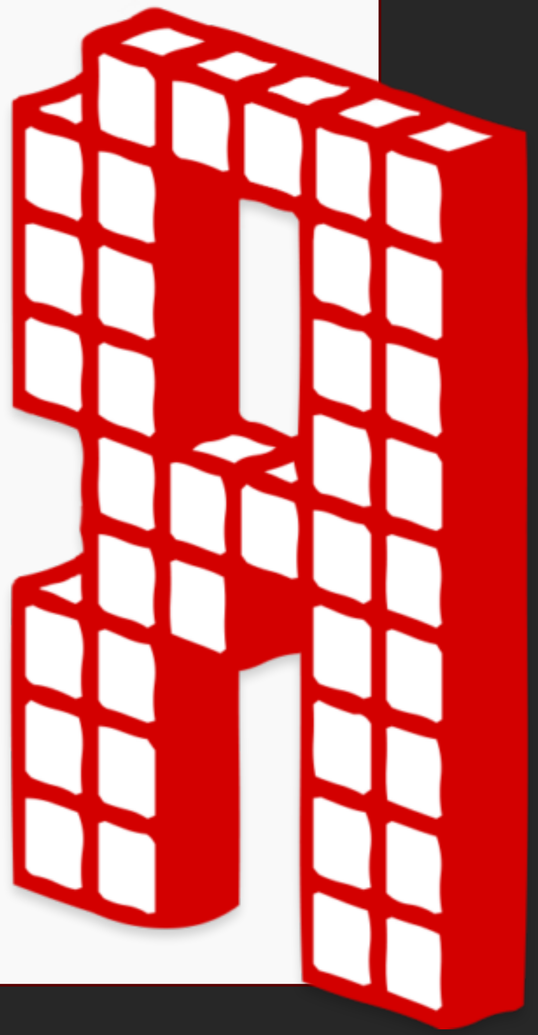
01.

TIDES

Truncation Induced DEpendencies among Summary statistics

Pacman plot



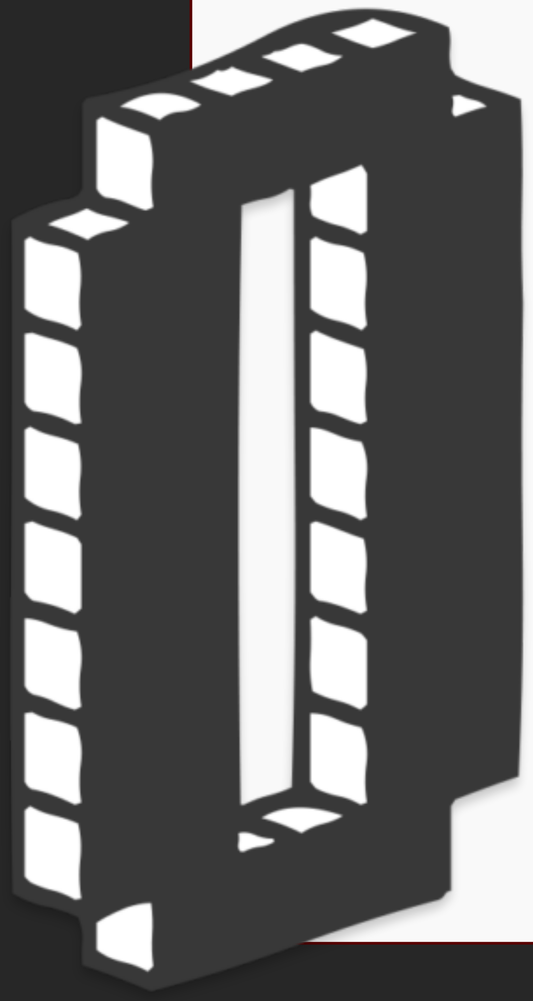


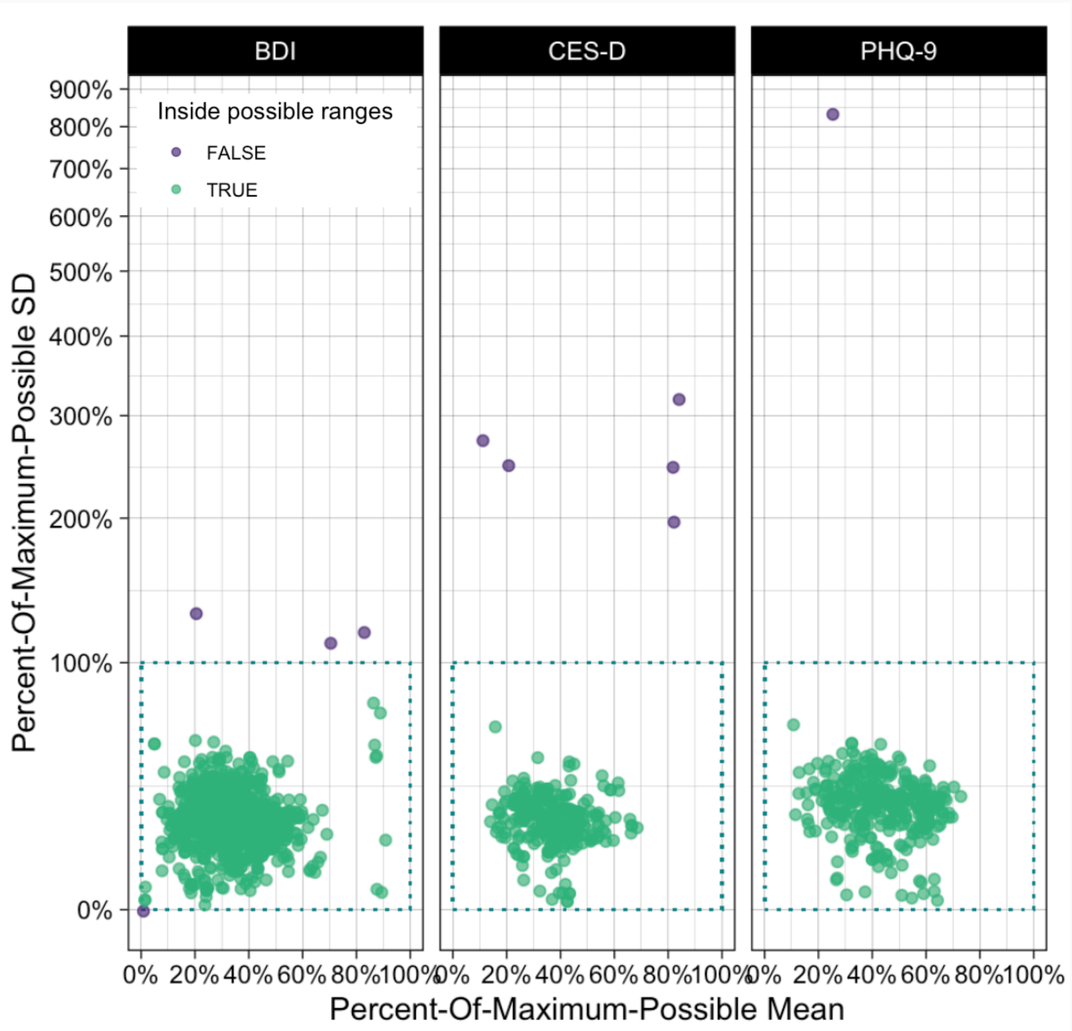
Hussey et al. (in prep)

Standardized via Percent-Of-Maximum-Possible

TIDES

Truncation Induced DEpendencies among Summary statistics





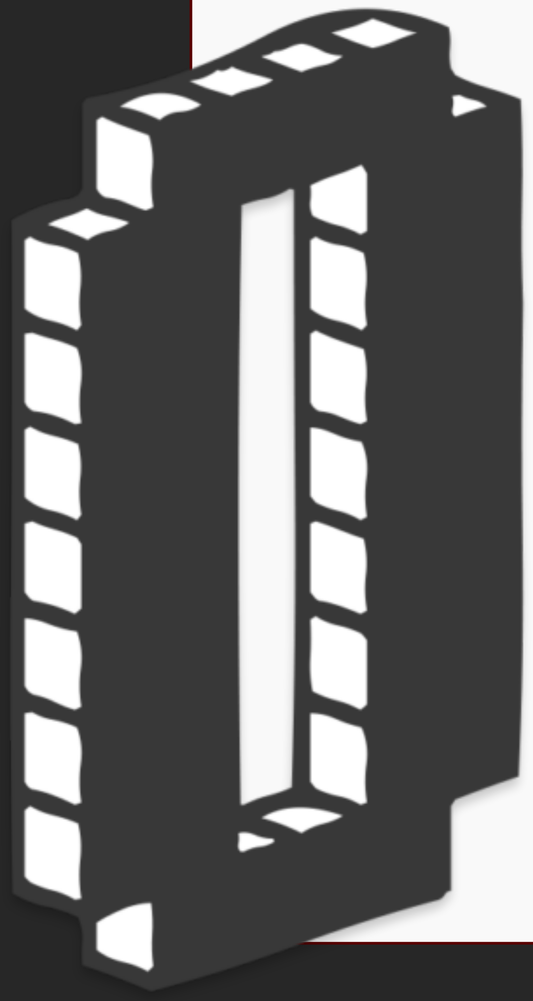
TIDES POMP plot

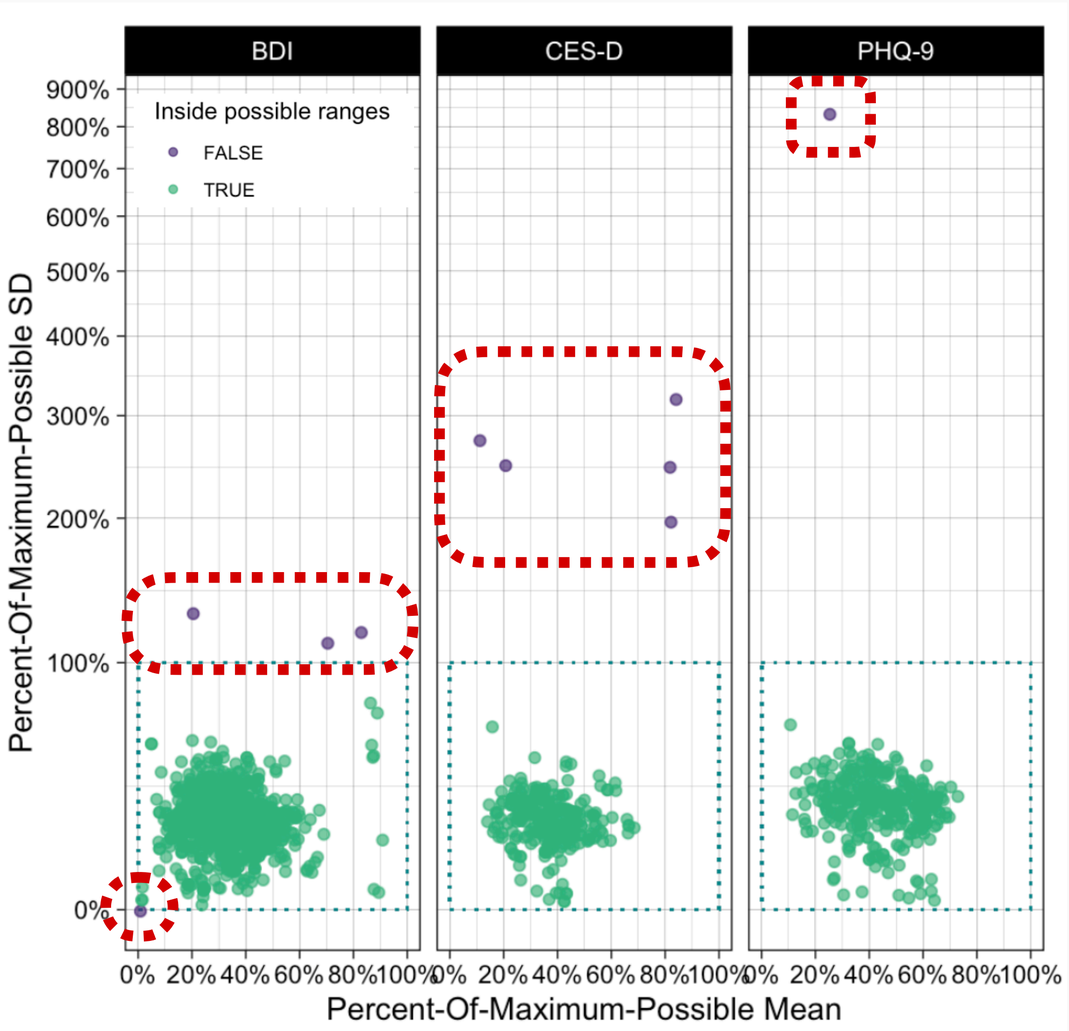
Aids general intuitions
Lowers necessary expertise

metapsy database of RCTs on psychotherapy for depression

TIDES

Truncation Induced DEpendencies among Summary statistics





TIDES POMP plot

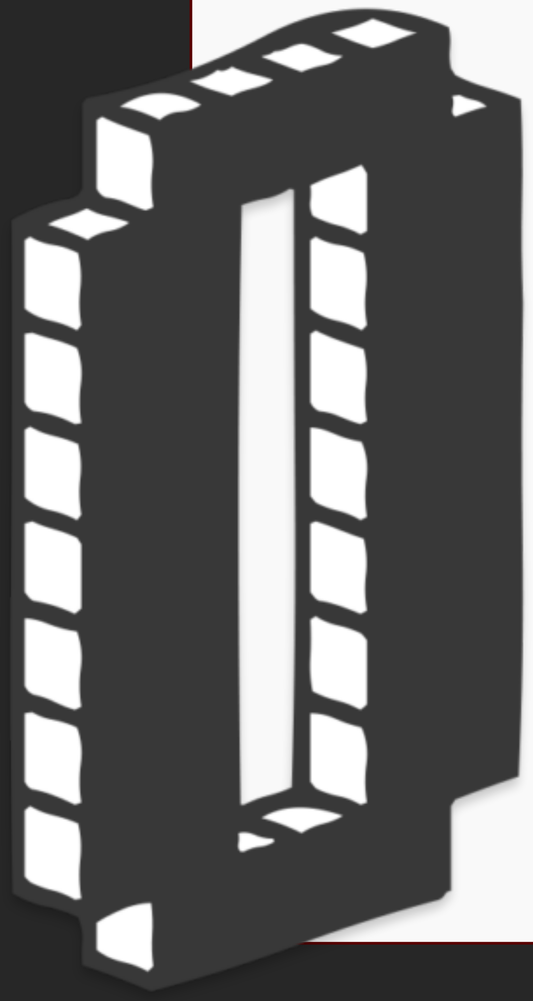
Aids general intuitions
Lowers necessary expertise

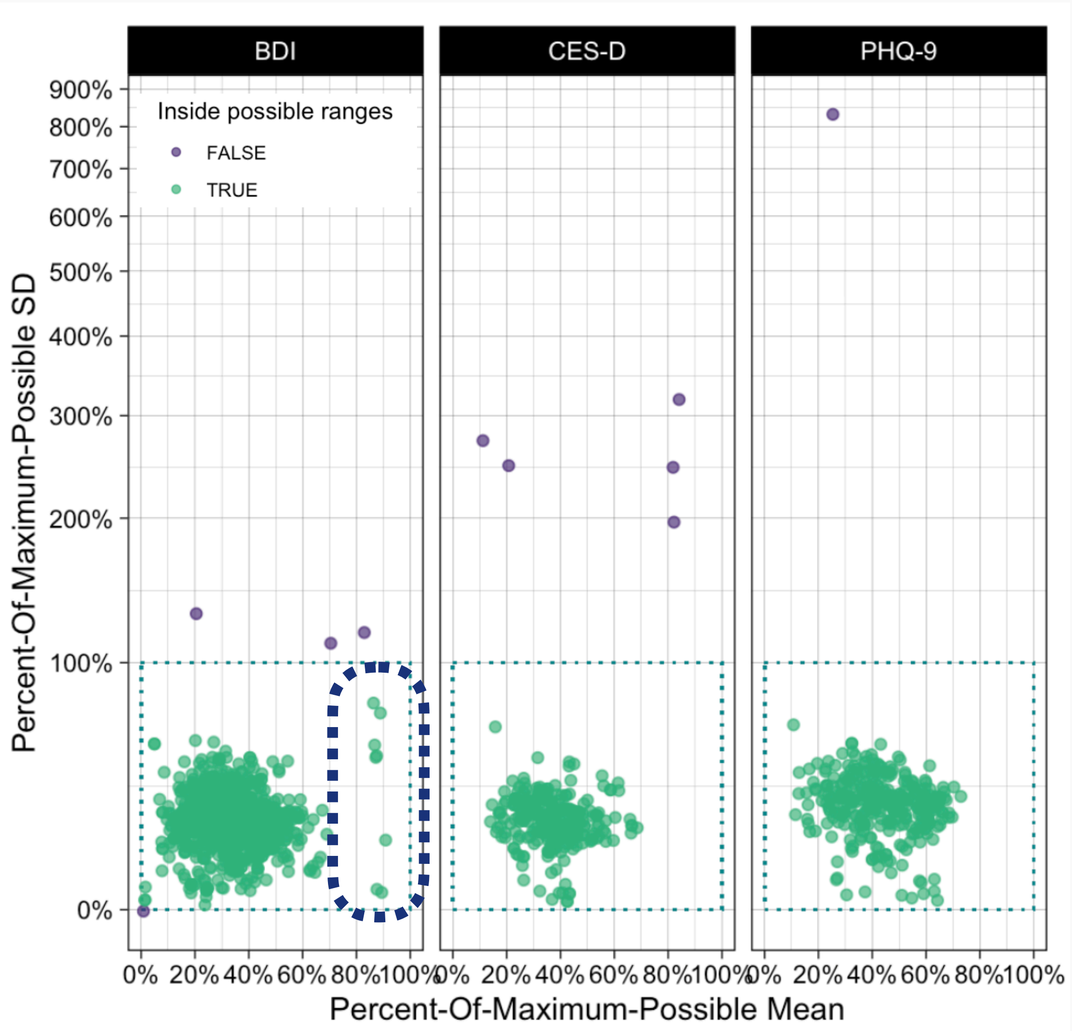
metapsy database of RCTs on psychotherapy for depression

- Some are impossible

TIDES

Truncation Induced DEpendencies among Summary statistics





TIDES POMP plot

Aids general intuitions
Lowers necessary expertise

metapsy database of RCTs on psychotherapy for depression

- Some are impossible
- Others are implausible

01.

Error / trustworthiness checks

Expertise problem

Beck Depression Inventory (BDI-II)

$N = 23, M = 20.70, SD = 3.40$

errors.shinyapps.io/tides

01.

Error / trustworthiness checks

Expertise problem

Beck Depression Inventory (BDI-II)

$N = 23, M = 20.70, SD = 3.40$

Patient Health Questionnaire 9 (PHQ-9)

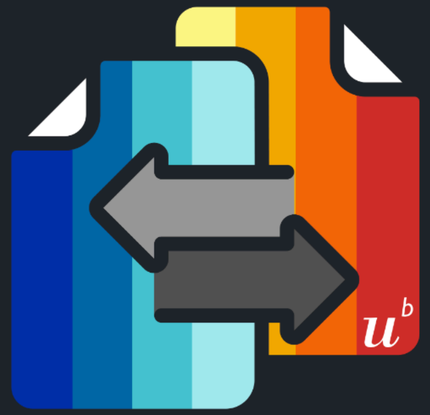
$N = 20, M = 14.50, SD = 14.13$

errors.shinyapps.io/tides

regcheck.app

RegCheck

Home Demo Tools Team Contact



Compare preregistrations with papers. Instantly.

RegCheck leverages large-language models and AI to automatically compare preregistration plans with scientific publications. This lets researchers effortlessly identify if and how executed studies deviated from the initial plan.

Try it now

Jamie Cummins, Psychology of Digitalisation, University of Bern, 2024

Cover template for Bootstrap, by @mdo.

regcheck.app

RegCheck

Home Demo Tools Team Contact

Compare preregistrations with papers. Instantly.

This app is an **alpha version**. It has not yet been extensively tested.
Due to token limits, Llama 3 will only work with short papers.
ChatGPT should work in most cases.
Processing files may take a minute or two.

Choose your model:

ChatGPT-4o

Does the paper have multiple experiments?

No

Preregistration:

Choose file

No file chosen

or supply OSF link

Supported files types:
.txt and .pdf

Paper:

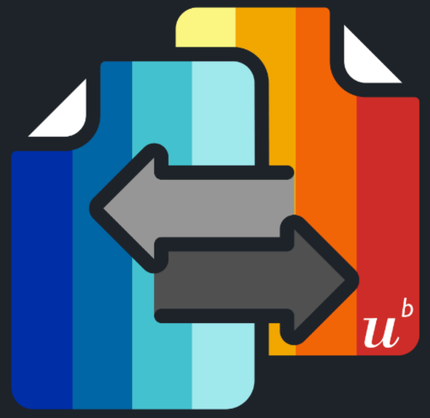
Choose file

No file chosen

or supply DOI

Compare

Study Feature	Information in Paper	Preregistered Protocol	Match
Final sample size	147 participants	150 participants	FALSE
Power Analysis Basis	95% power for medium effect size (f2 = 0.15)	No power analysis for first analysis; 95% power for medium effect size (f2 = 0.15) in second analysis	TRUE
Sampling Strategy	Recruit 150 and exclude those with incomplete data until sufficient sample size is achieved	Recruit 150 and apply exclusion criteria, then add in batches of 10 until at least 150 participants meet criteria	TRUE
Within-subjects factor	Prime Valence: positive vs. negative	Prime Valence: positive vs. negative	TRUE
Dependent Variables	Influence-awareness ratings, Target stimulus evaluations	Evaluations within the AMP as pleasant or unpleasant	FALSE
Exclusion criteria	Incomplete data	Completion time < 3 minutes, Partial data on demographics/AMP	TRUE
IA-AMP Task Trials	10 practice trials, 120 critical trials	10 practice trials, 120 main trials	TRUE
Exploratory	Three exploratory	Five exploratory questions	FALSE



In progress:

Validation against human-coded data

API support for OSF and clinicaltrials.gov

UNIVERSITÄT BERN

A master’s degree course in error detection

xxxxxxxxxxxxxxxxxxxxxxxxxx

02.

Conspicuous absence of error detection in the curriculum

Defence against the dark arts:
a proposal for a new MSc course
(Bishop, 2023)

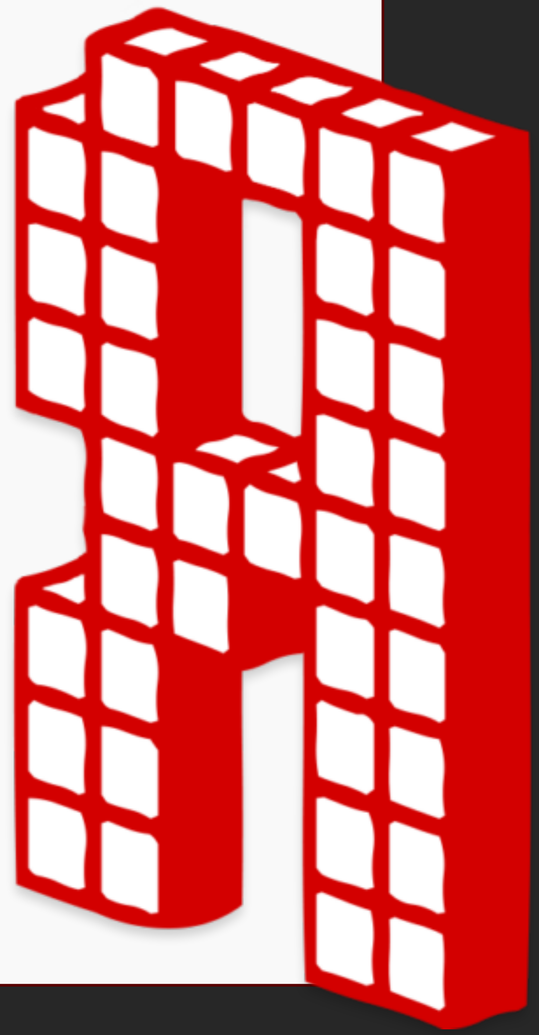
02.

“Estimating the credibility of past research”

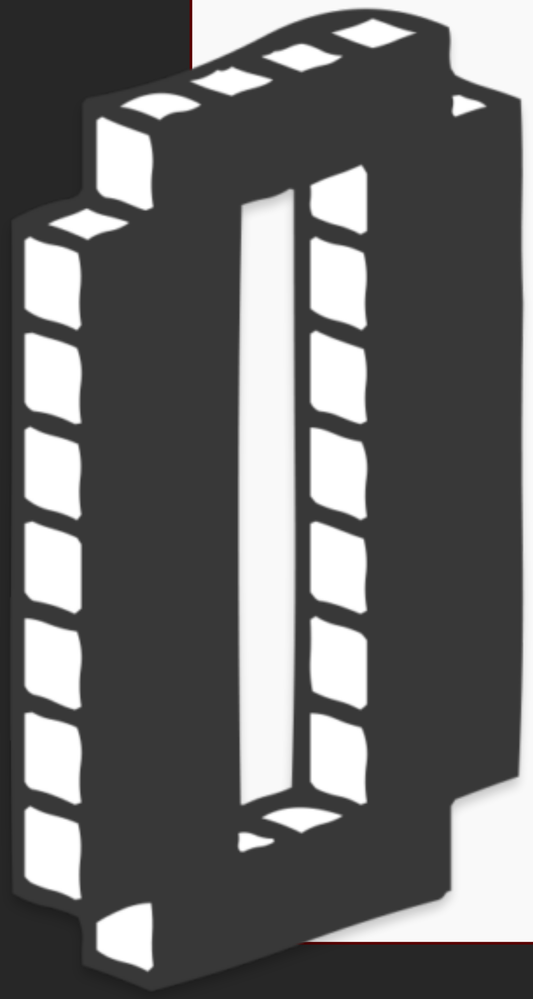
14 week Masters course

Running at University of Bern since Fall ’23

Science as an intensely human & fallible activity
Organised skepticism



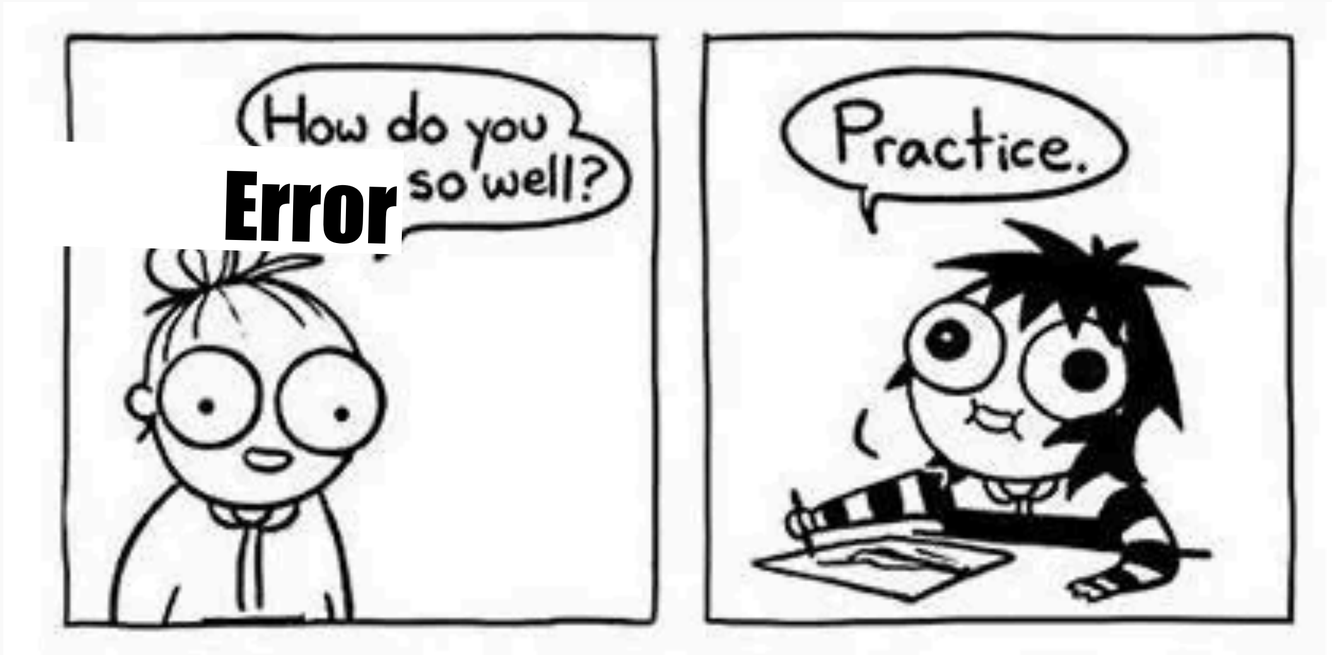
Syllabus



- Functions of science other than truth discovery
- Existing quality assurance mechanisms in science and their flaws
- StatCheck, GRIM, GRIMMER, TIDES, SPRITE, etc
- Effect size plausibility
- (Missing) causal language
- Explanans/explanandum confusion
- Jingle/jangle fallacies and measurement issues
- Obtaining & extracting data
- Organise and communicate critiques (INSPECT-SR, PubPeer)

Open Source materials will be released this spring!

02.



Error

Permission to critique

xxxxxxxxxx

Language of critique

UNIVERSITÄT BERN

ERЯOR

A bug bounty program for science

xxxxxxxxxxxxxxxxxxxxxxxxxx

03.

UNIVERSITÄT BERN

ADVI2ORY BOARD

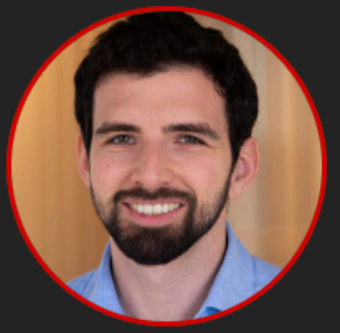
Dorothy Bishop
University of Oxford

Nick Brown
Linnaeus University

Matthias Egger
University of Bern
University of Cape Town
University of Bristol

Julia Rohrer
University of Leipzig

Anne Scheel
Utrecht University



Leo Tiokhin
Eindhoven University of Technology

Richard McElreath
Max Planck Institute for Evolutionary Anthropology

Brian Nosek
Center for Open Science

Michèle Nuijten
Tilburg University

Simine Vazire
University of Melbourne

250,000 CHF

(€260k, $285k)

fund

Check

100

published articles
for errors

Pay authors & reviewers

+ bonus rewards

contingent on errors found

+ offer co-authorship of umbrella publications

Goals

- Explore and test **practical challenges** in the implementation of an error checking system

- Obtain empirical estimates of the
the **prevalence** of errors and their **types**
  - cf. OSC (2015) Estimating the Reproducibility of Psychological Science

- **Intervention** on our scientific culture around errors
  - Providing positive examples of scientists self-correcting
  - Specify reinforcers for error acceptance
  - Extinguish perceived punishers

Candidate articles

Obtain author consent

Match with reviewer

- Selected for ‘importance’
  Would matter if it contained errors.
  30+ citations per year.
  Recent (c.10 years)

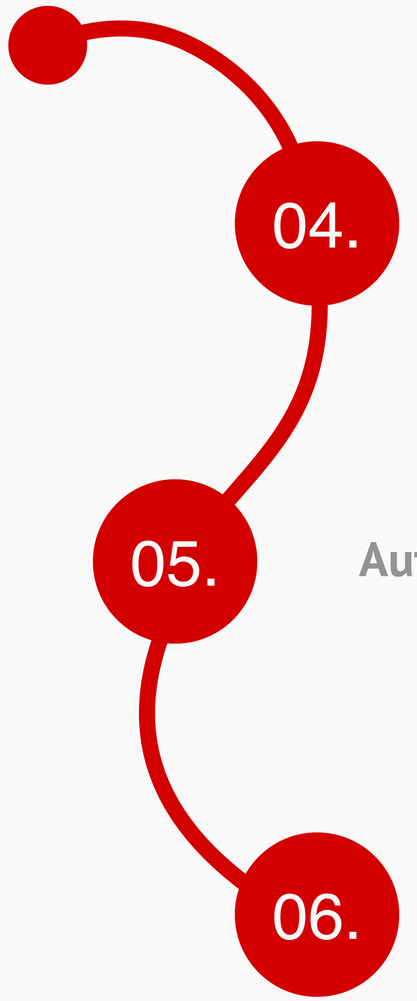
- **Not** randomly sampled

- Attention paid to gender, seniority, etc.

- **Consensual** process

- Pre-commit to:
  - Sharing materials
  - Answering reasonable questions
  - **Acting** on recommendations

- **Not** anonymous
- Can ask author reasonable questions to aid their work



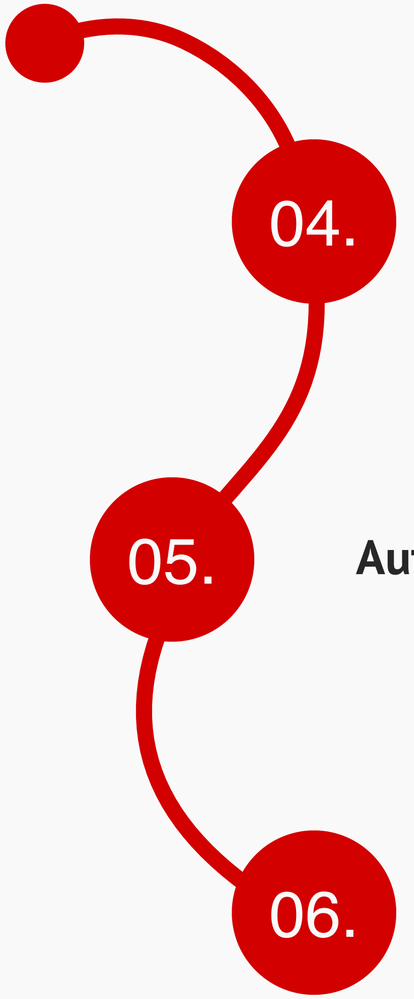
Review

Author response

ERROR Report

- Checked by Recommender

- Claims must be substantiated

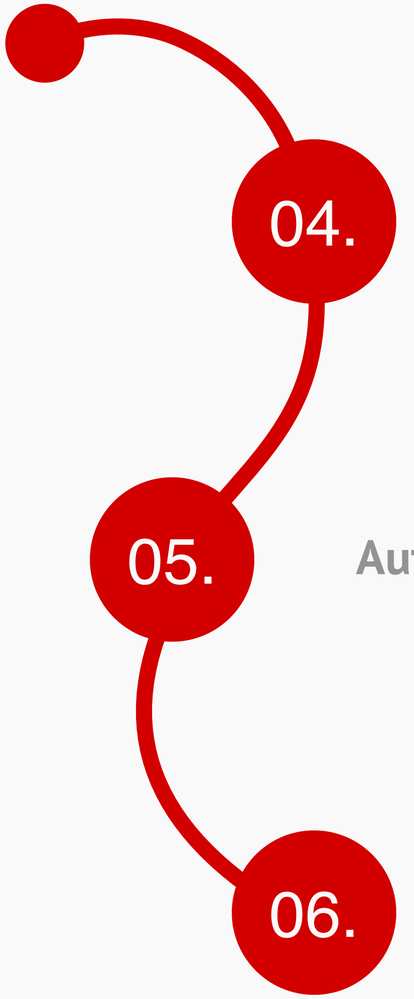


Review

Author response

ERROR Report

- **Optional**, timely
- Checked by Recommender
- Claims must be substantiated



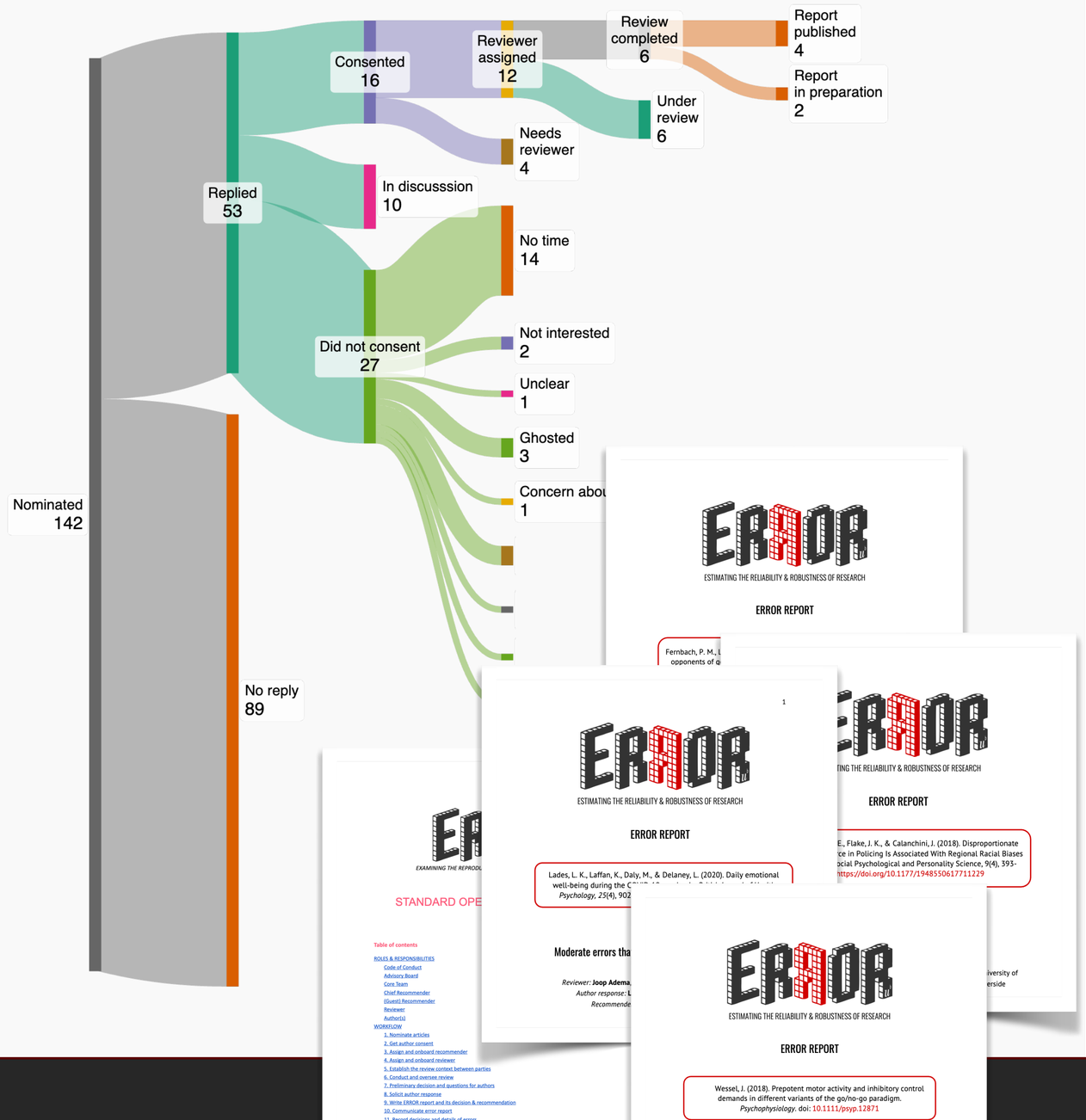
- Recommender makes a **Decision** & associated **Recommendation(s)**

- Published
  - error.reviews website
  - PsyArXiv
  - PubPeer
  - CC-By 4.0 licence

Decision	**Recommendation**	**Bonus reward** (indicative)
No errors	No additional action beyond publication of error report on the error.reviews website.	250 CHF to the author
Minor error(s)	Authors to appropriately recognise these errors in future discussions of the article. *Probably most research!*	250 CHF to both the author & reviewer
Indeterminable error(s)	No determination could be made re the presence or absence of important potential errors. *Less desirable than verifiably minor errors!* Authors to appropriately recognise this lack of verifiability in future discussions of the article.	250 CHF to the reviewer
Moderate error(s)	Correction notice (minor)	500 CHF to the reviewer
Major error(s)	Correction notice (major) / may warrant an expression of concern	1000 CHF to the reviewer
Severe error(s)	Retraction	2500 CHF to the reviewer

+ error(s) do / do not affect core conclusions

Snapshot of activities at 1 year



THE CHRONICLE OF HIGHER EDUCATION

Sign In

RESEARCH INTEGRITY

Wanted: Scientific Errors. Cash Reward.

By Stephanie M. Lee | February 21, 2024

---

nature

Explore content | About the journal | Publish with us | Subscribe

nature > technology features > article

TECHNOLOGY FEATURE | 19 August 2024

Cash for errors: project offers bounty for spotting mistakes in published papers

ERROR project borrows a strategy from the software industry.

By Julian Nowogrodzki

RESEARCH INTEGRITY

Wanted: scientific bounty hunters

A pilot program will pay reviewers to check important published papers and preprints in psychology. Reviewers will receive up to 3500 Swiss francs (nearly $4000) depending on the severity of any errors uncovered. Authors must consent in advance; they, too, will receive compensation if they cooperate and their work proves reliable. The program, Estimating the Reliability and Robustness of Research, is funded by the University of Bern and modeled after payments by software companies to programmers who find flaws in code. Backers say scientific authors and reviewers lack incentives to identify errors in the literature.

WIRED SECURITY POLITICS GEAR THE BIG STORY BUSINESS SCIENCE CULTURE IDEAS MERCH SIGN IN

MATT REYNOLDS SCIENCE JUN 21, 2024 7:00 AM

Science Is Full of Errors. Bounty Hunters Are Here to Find Them

A new project is paying researchers to find errors in other scientists' work. The only problem? Even error hunters make mistakes.

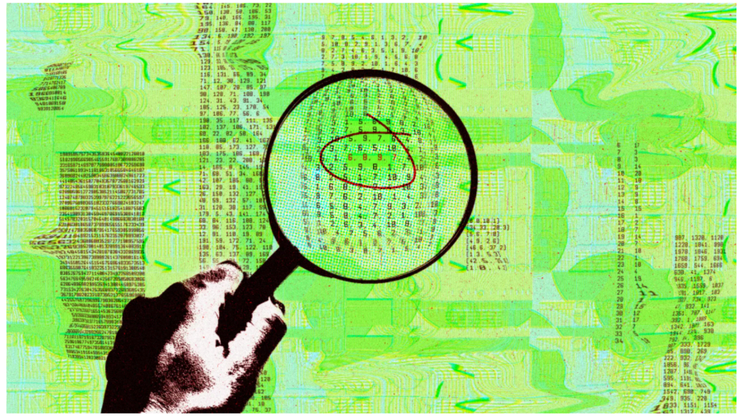


PHOTO-ILLUSTRATION: ROSIE STRUVE; GETTY IMAGES

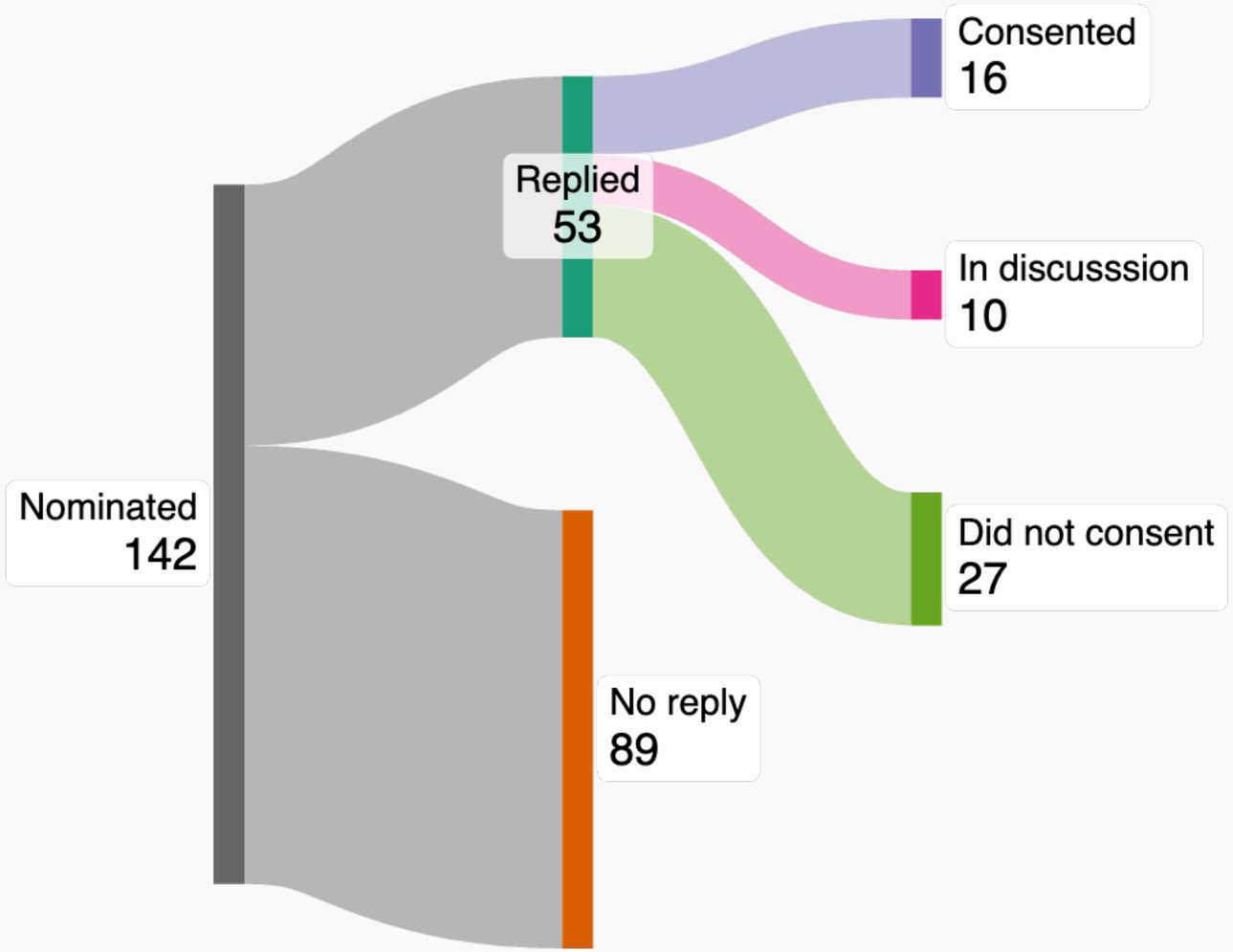
THE TRANSMITTER

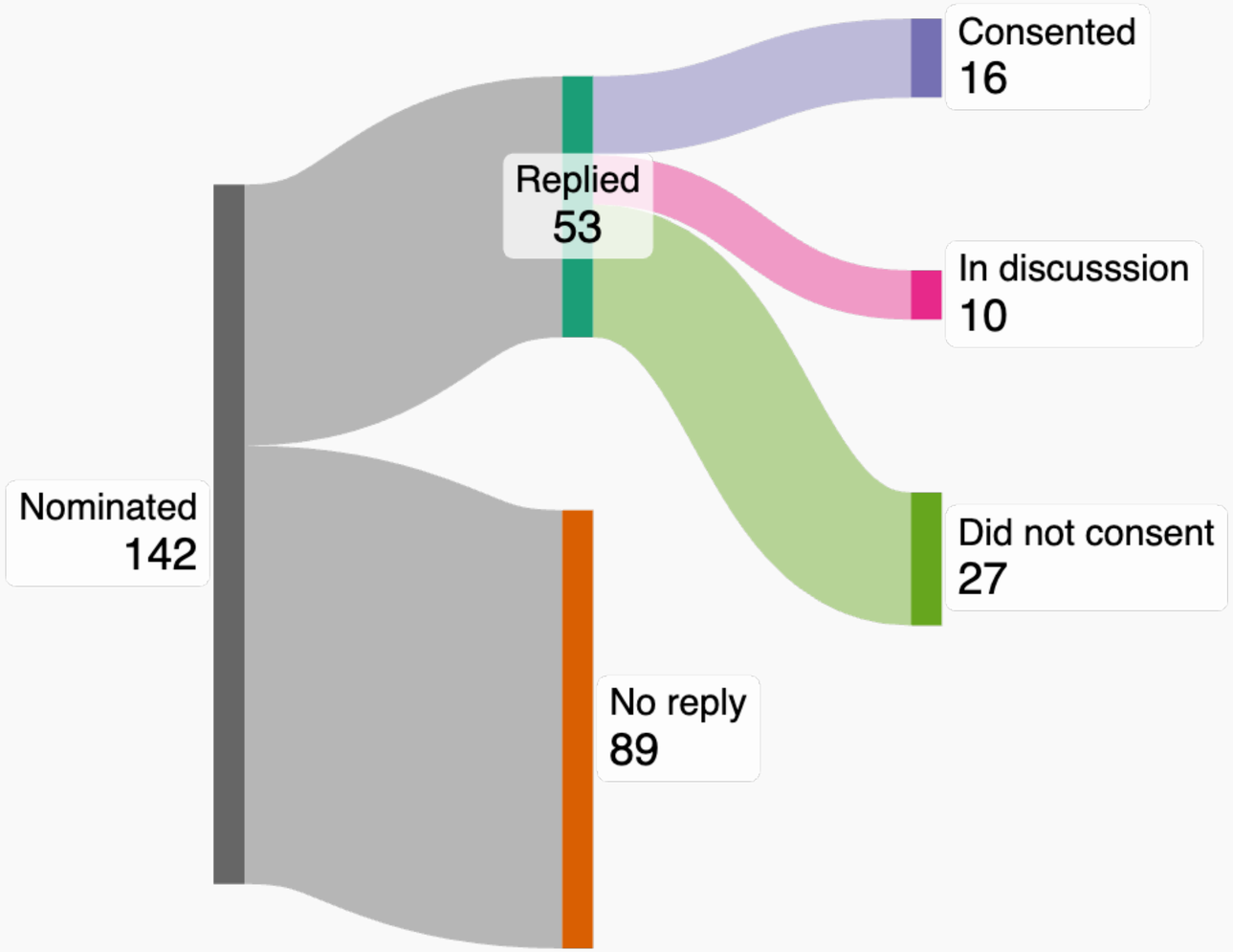
NEWS / PUBLISHING

Bounty hunting for blunders: Q&A with Russell Poldrack and Jan Wessel

The guinea pigs for a post-publication error-spotting project discuss why the field should destigmatize slipups—and how to handle them better.

BY CALLI MCMURRAY
7 JUNE 2024 | 10 MIN READ


Nominated
142
Replied
53
No reply
89
Consented
16
In discusssion
10
Did not consent
27



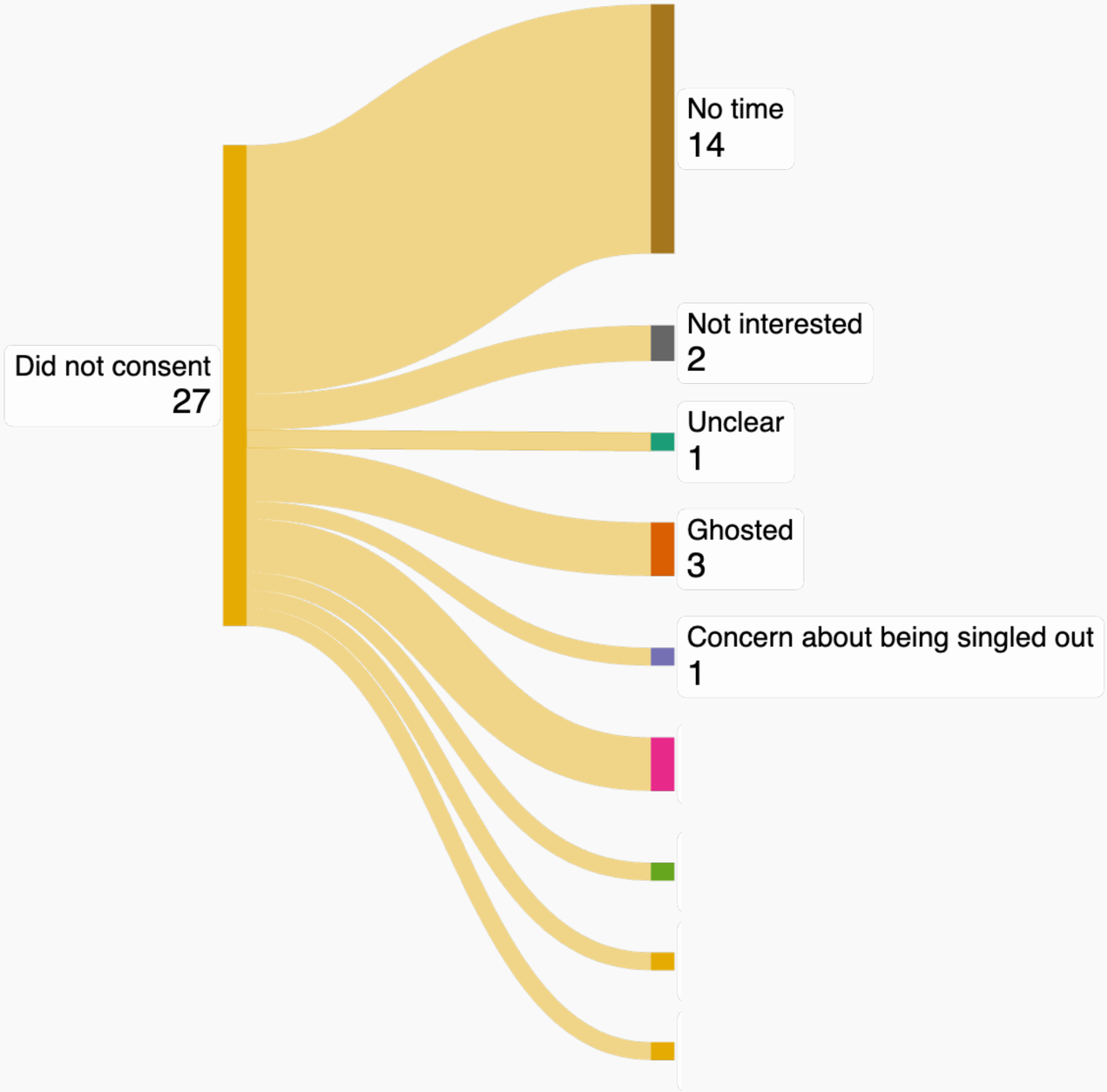
Response rate

37%

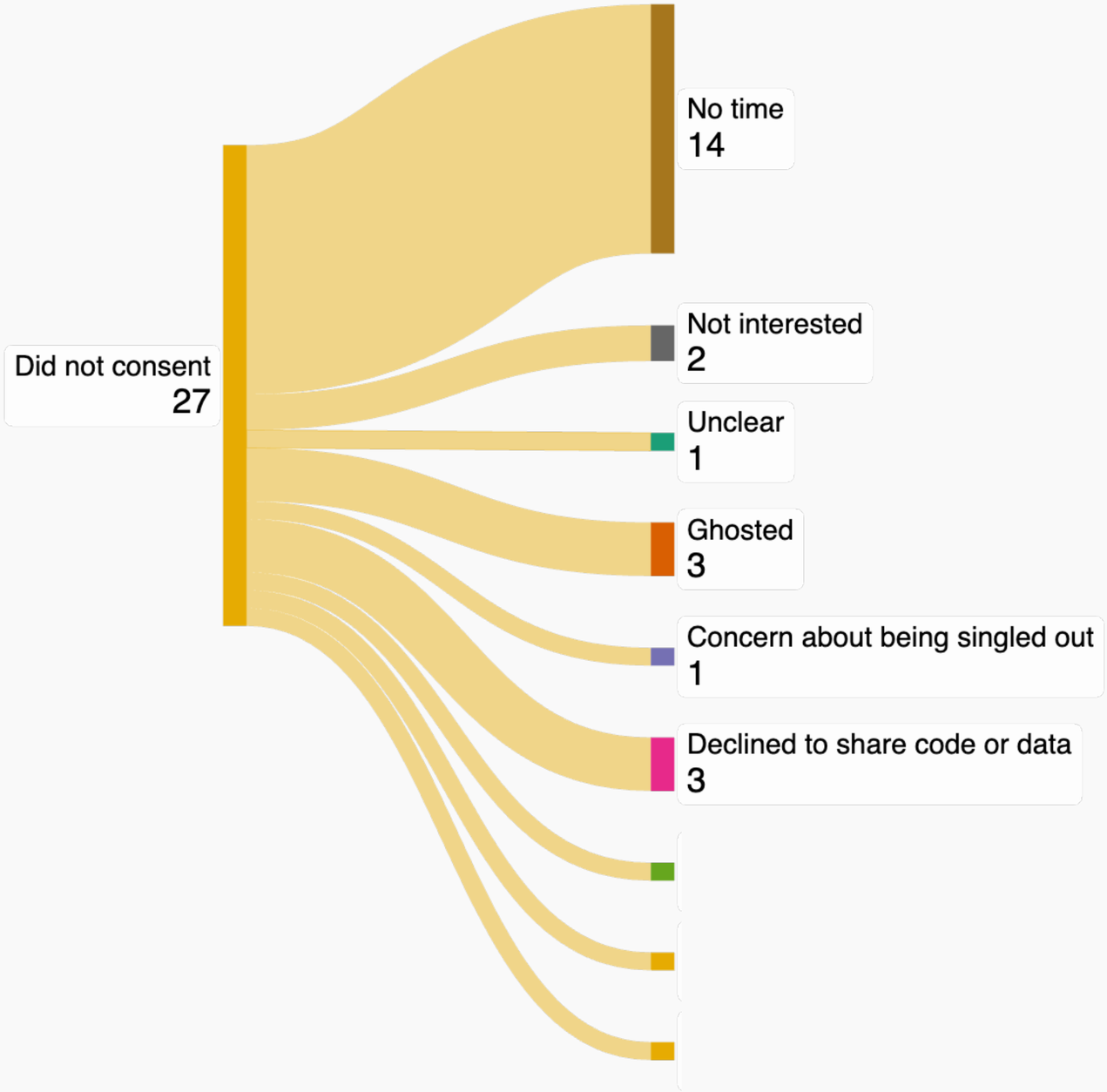
Consent rate

11%

(30% of responders)

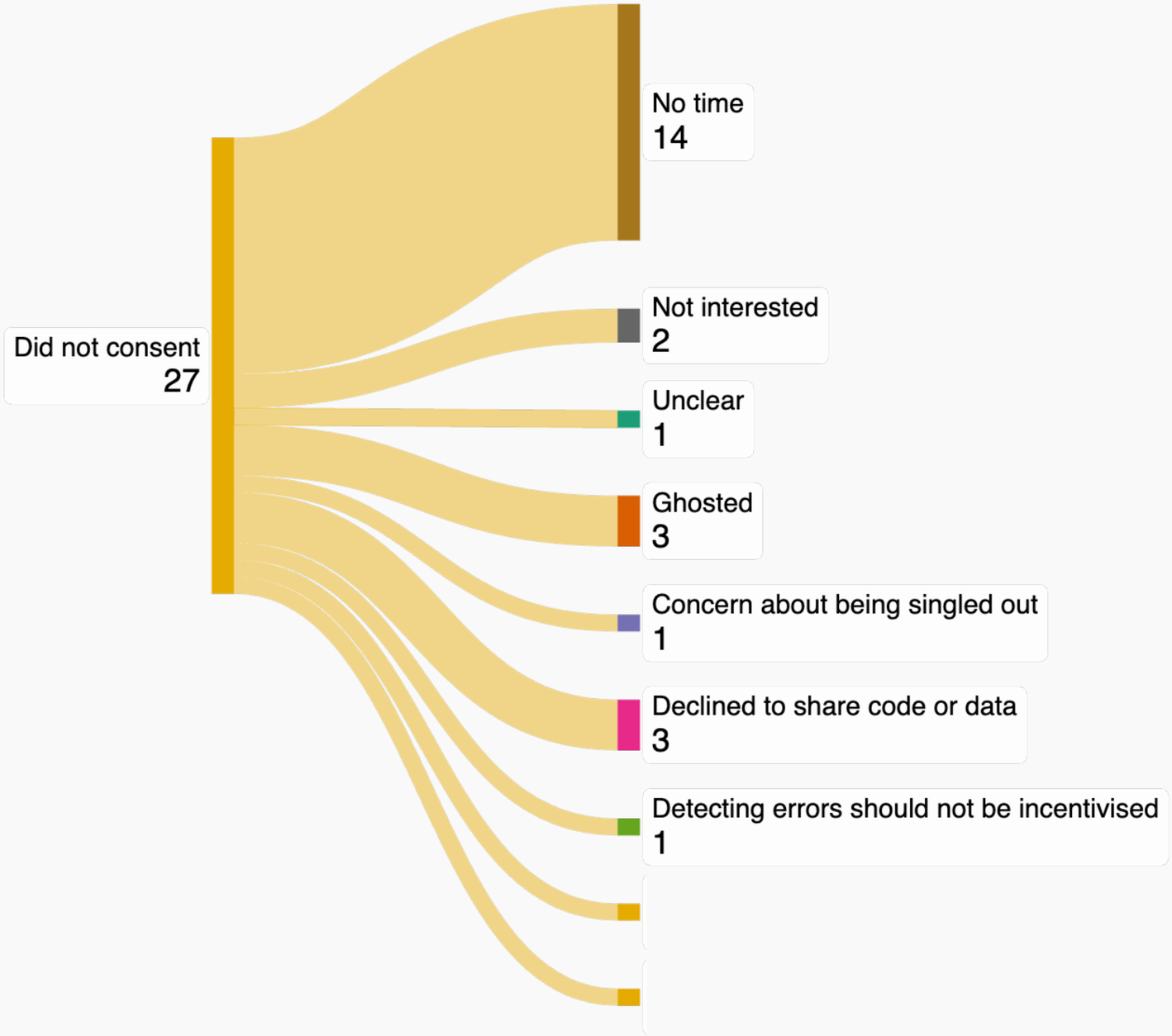


Reasons
for non-consent

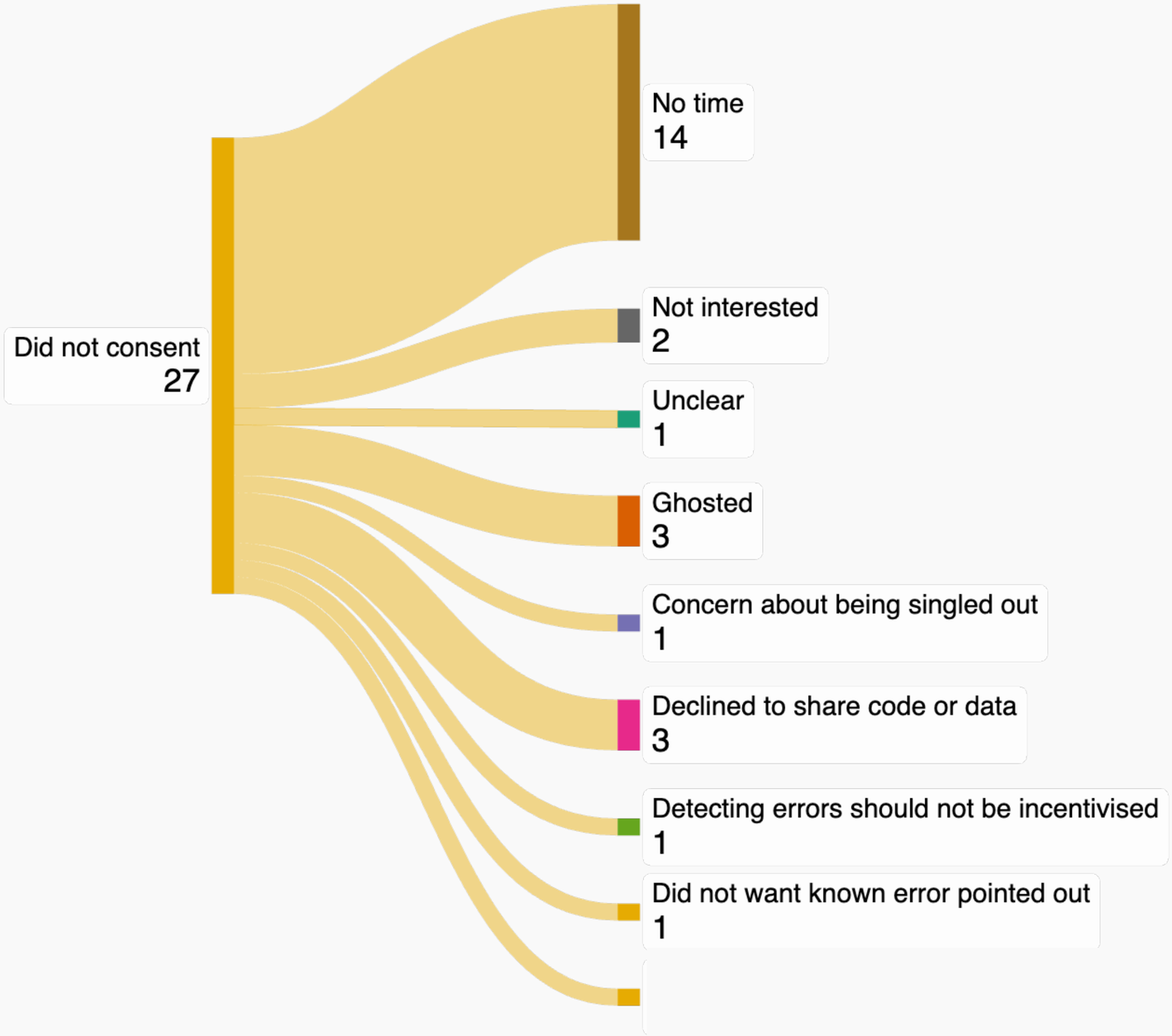


Reasons
for non-consent

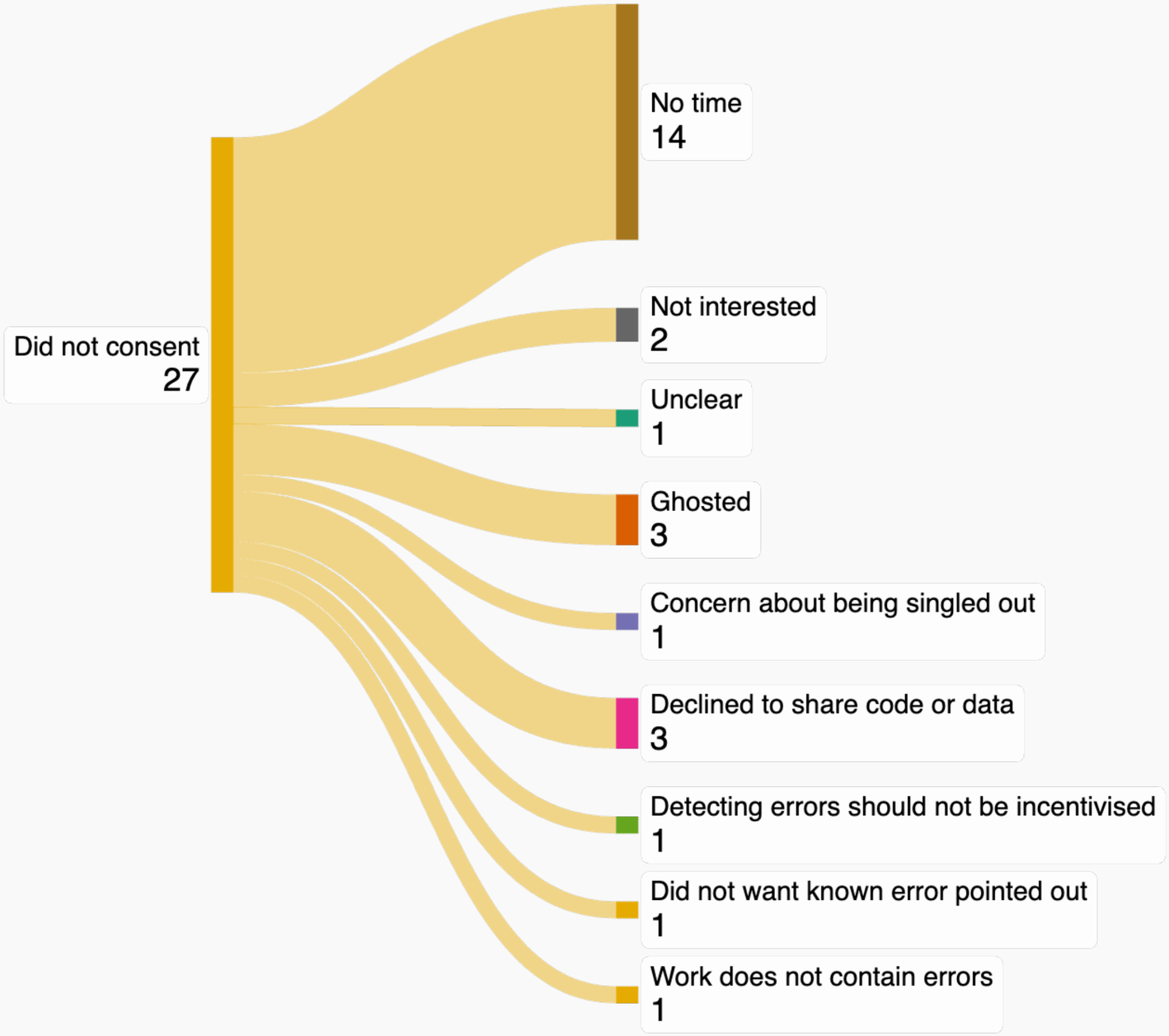
Reasons
for non-consent

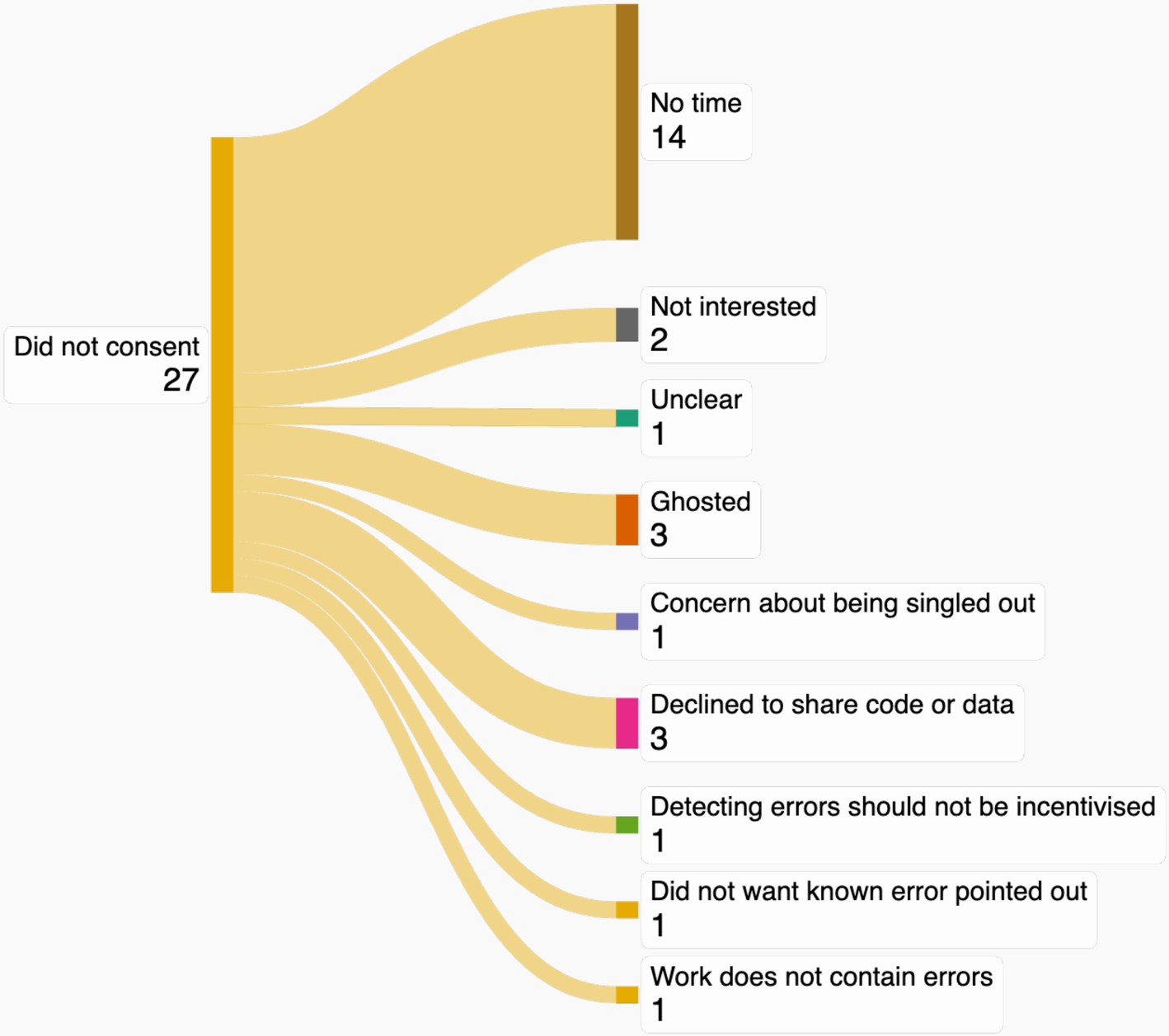


Reasons
for non-consent



Reasons
for non-consent

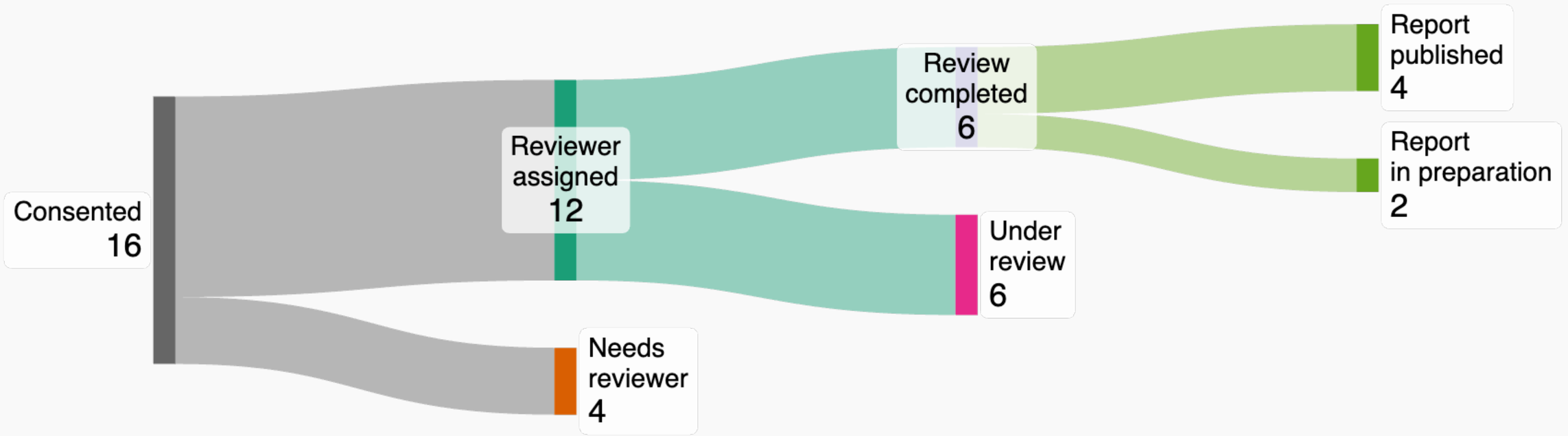




Reasons
for non-consent

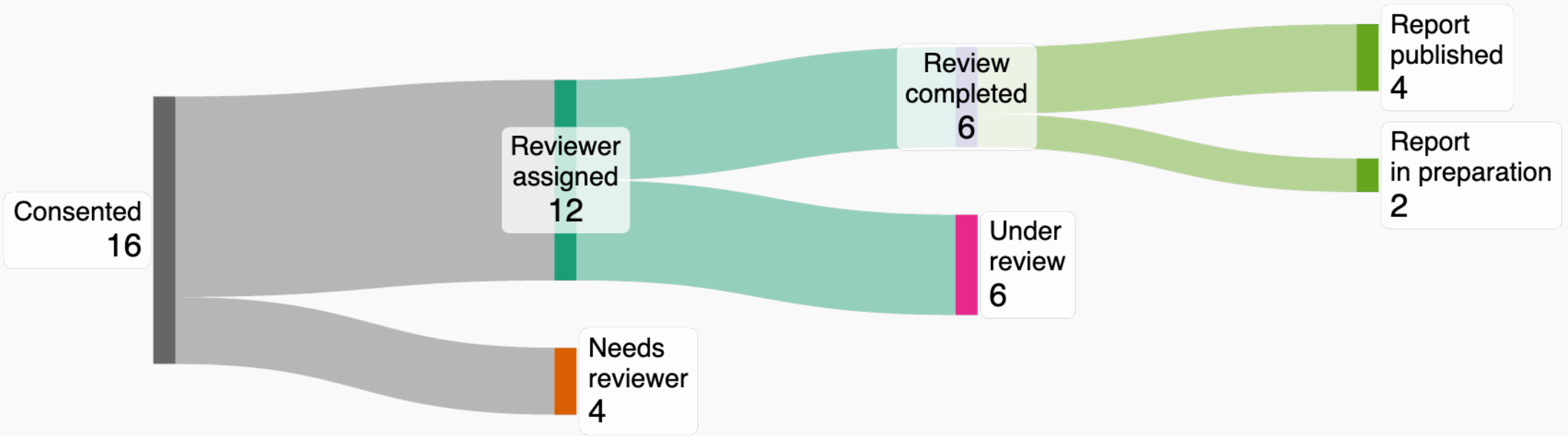
Reasons we find difficult to defend:

22%
of responders
(4% of nominated)



Difficulty finding reviewer or completing review

25%



Building momentum

- Guest Recommenders
- Explicated processes

4

Reviews Completed

Wessel, J. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*. doi: 10.1111/psyp.12871

Lades, L. K., Laffan, K., Daly, M., & Delaney, L. (2020). Daily emotional well-being during the COVID-19 pandemic. *British Journal of Health Psychology, 25*(4), 902-911. https://doi.org/10.1111/bjhp.12450

Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y., & Rozin, P. (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour, 3*(3), 251-256. https://doi.org/10.1038/s41562-018-0520-3

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents. Social Psychological and Personality Science, 9(4), 393-401. https://doi.org/10.1177/1948550617711229

4

Reviews
Completed

4

Papers
with Errors

Decisions & Recommendations

- **Minor** Errors that do **not** affect the core conclusions
  2 articles

- **Moderate** Errors that do **not** affect the core conclusions
  1 article

- **Major** Errors that **affect** the core conclusions
  1 article

Decisions & Recommendations

- **Minor** Errors that do **not** affect the core conclusions
  2 articles
  - No correction recommended, but 1 author seeks one anyway

- **Moderate** Errors that do **not** affect the core conclusions
  1 article
  - Correction recommended, but unclear if authors will act

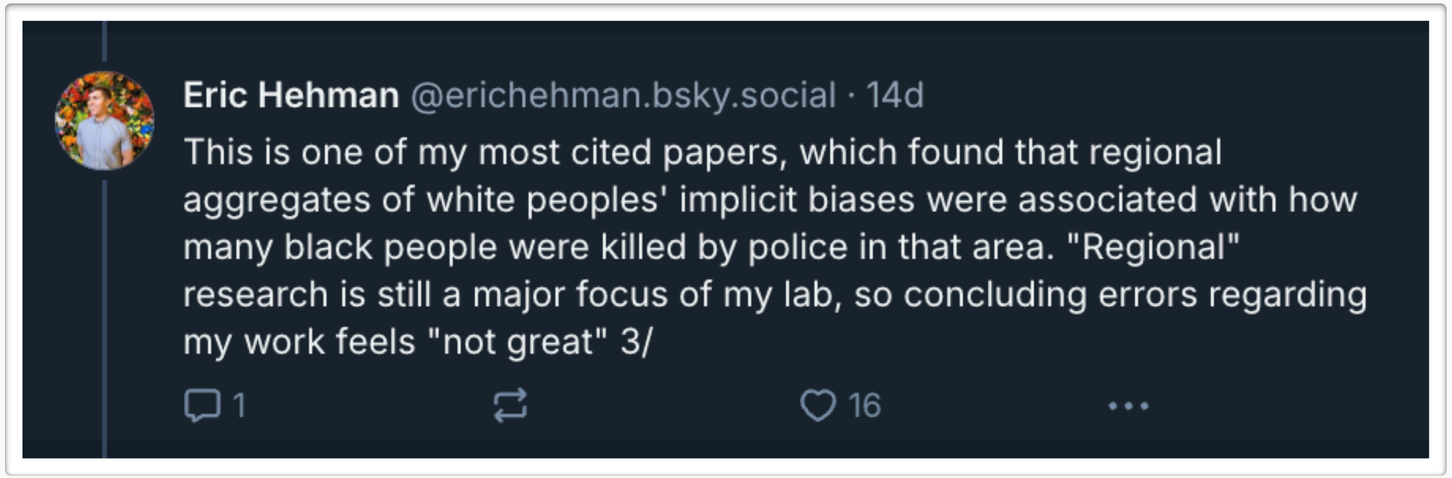
- **Major** Errors that **affect** the core conclusions
  1 article
  - Correction recommended, but unclear if authors will act

Common Errors Encountered

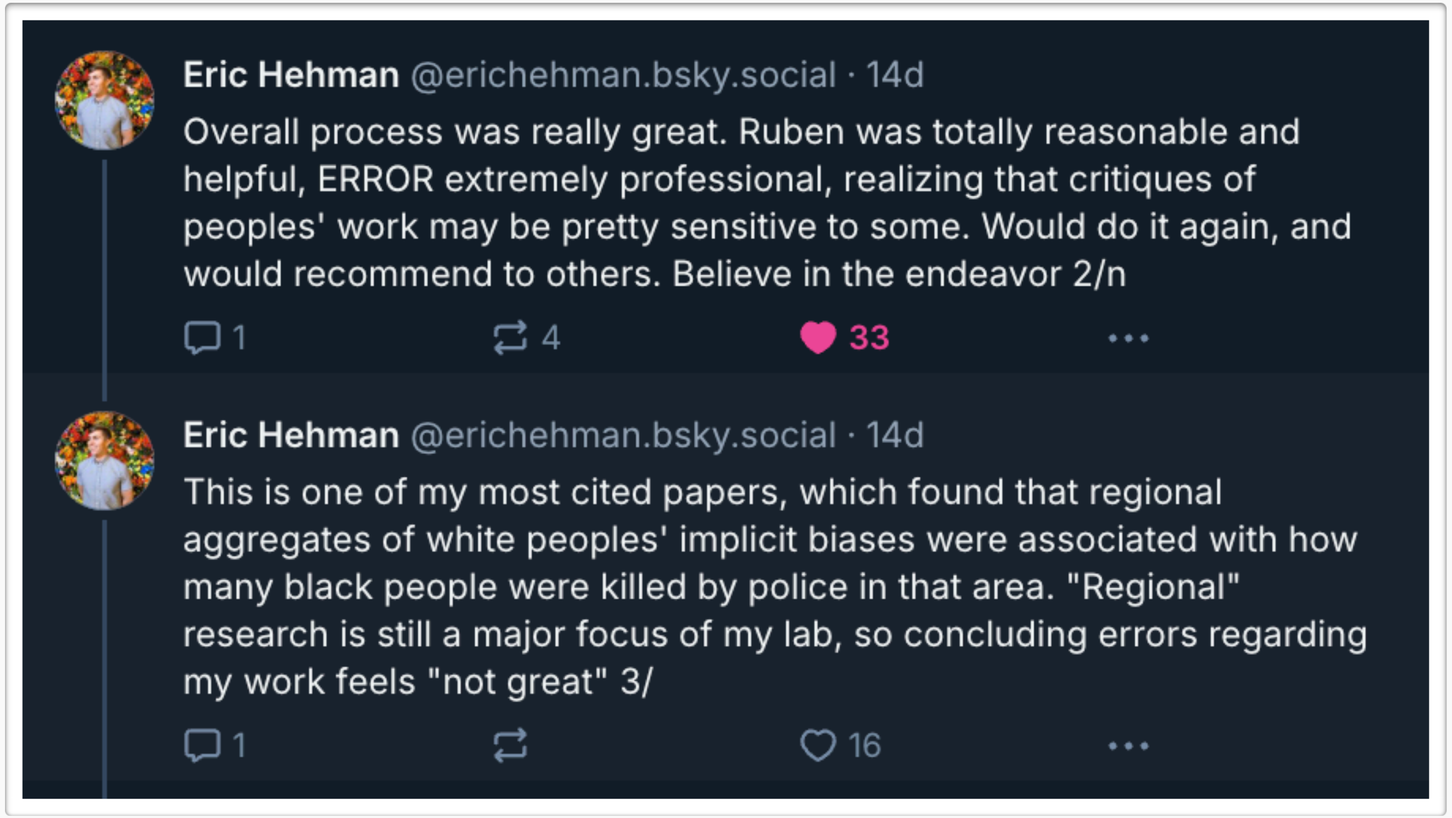
[very preliminary]

- Non-significant *p* values treated as evidence for the null
- Poor reproducibility in early stages of data cleaning/processing
- Informally: Prevalence of not-even-wrong claims

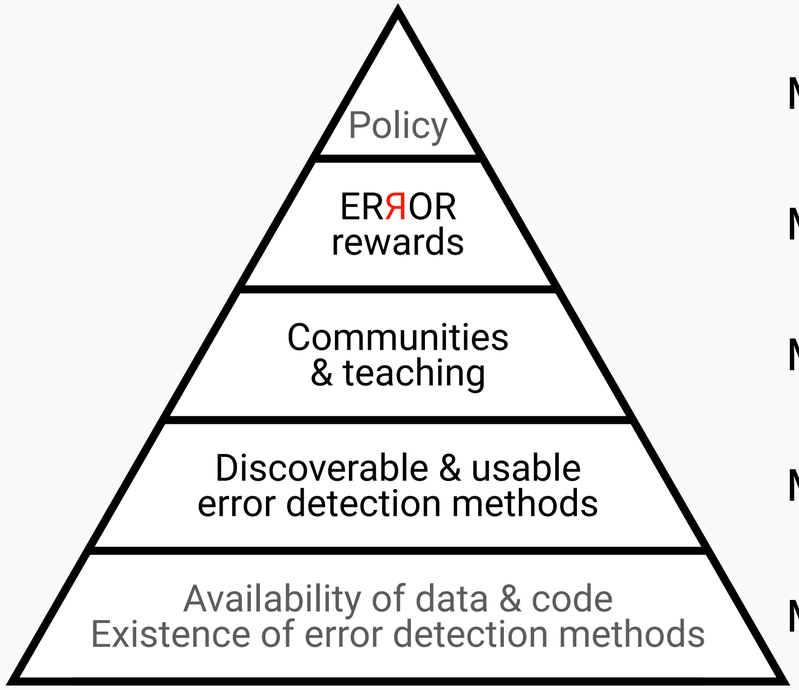
Experiences & Reactions



Experiences & Reactions



Our strategy to increase error checking, reporting, & correction



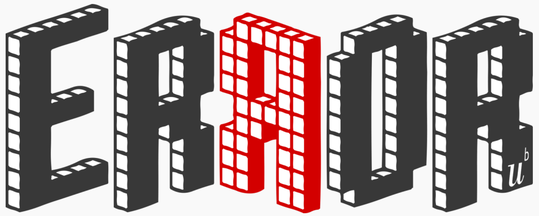
Make it required

Make it rewarding

Make it normative

Make it easy

Make it possible



Contact

ian.hussey@unibe.ch
Bluesky: @ianhussey.bsky.social

https://error.reviews
error-reviews.psy@unibe.ch
Bluesky: @error.reviews

UNIVERSITÄT
BERN

xxxxxxxxxxxxxxxxx

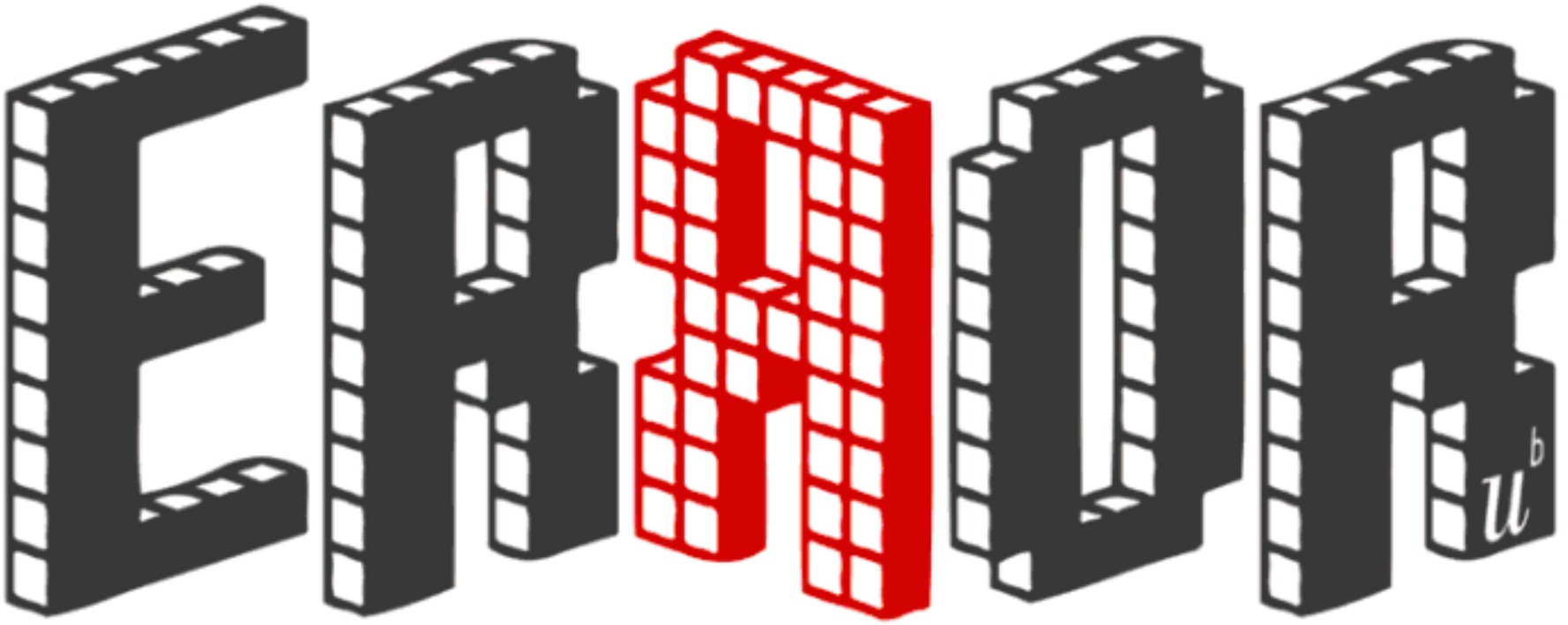
Many thanks to

Collaborators
- Ruben Arslan
- Jamie Cummins
- Malte Elson

ERROR Advisory Board

Authors & Reviewers
- Joop Adema
- Jimmy Calanchini
- Jessica K. Flake
- R. Chris Fraley
- Eric Hehman
- Leonhard Lades
- Nick Light
- Will Lowe
- Russ Poldrack
- Jan Wessel

xxxxxxxxxxxxxxxxx



ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH