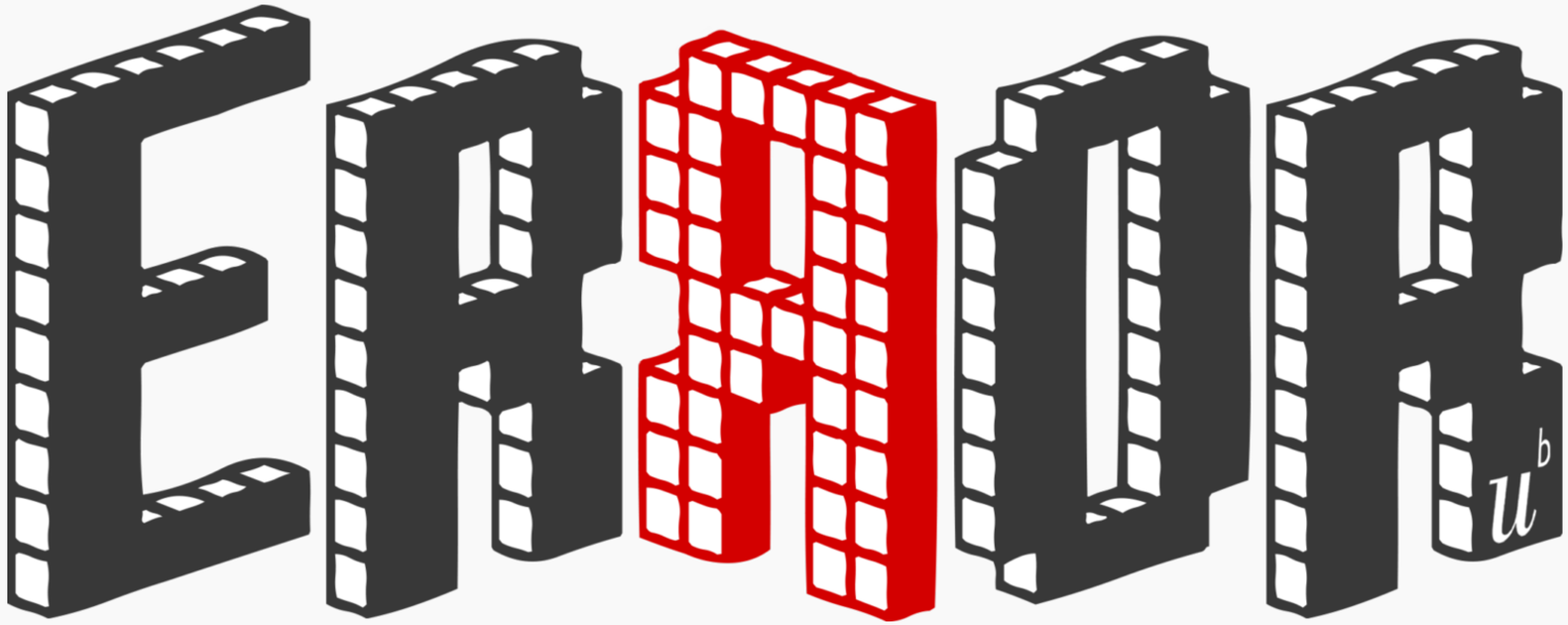
UNIVERSITÄT BERN

xxxxxxxxxxxxxxxxxxxxxxxxxxxx



xxxxxxxxxxxxxxxxxxxxxxxxxxxx

A three-pronged effort to improve post-publication critique & error detection

Ian Hussey

UNIVERSITÄT BERN

PRIИCIPAL INVESTIGATORS

Malte Elson
University of Bern

Ruben Arslan
University of Leipzig

CHIEF ЯECOMMENDER

Ian Hussey
University of Bern

UNIVERSITÄT BERN

ADVI2ORY BOARD

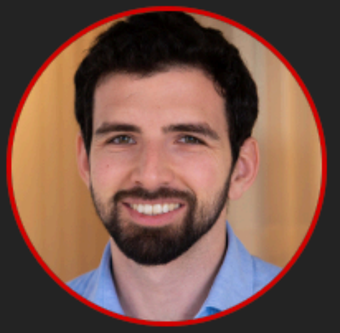
Dorothy Bishop
University of Oxford

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

xxxxxxxxxxxxxxxxxxxxxxx

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

xxxxxxxxxxxxxxxxxxxxxxx

Who can help carry this burden?

Need for parallel systems of scientific verification

Who can help carry this burden?

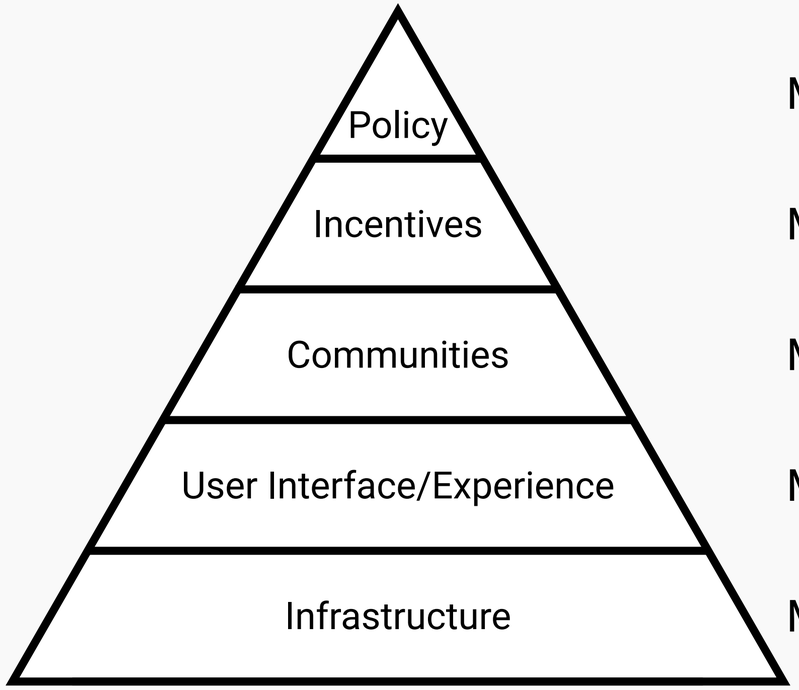
A clone-army of Nick Browns, Dorothy Bishops, & Elisabeth Biks

Who can help carry this burden?

"A few outsiders and weirdos saw the giant lie at the heart of [science], and they saw it by doing something the rest never thought to do:

They looked."

Centre for Open Science's strategy to increase preregistration & data sharing



Make it required

Make it rewarding

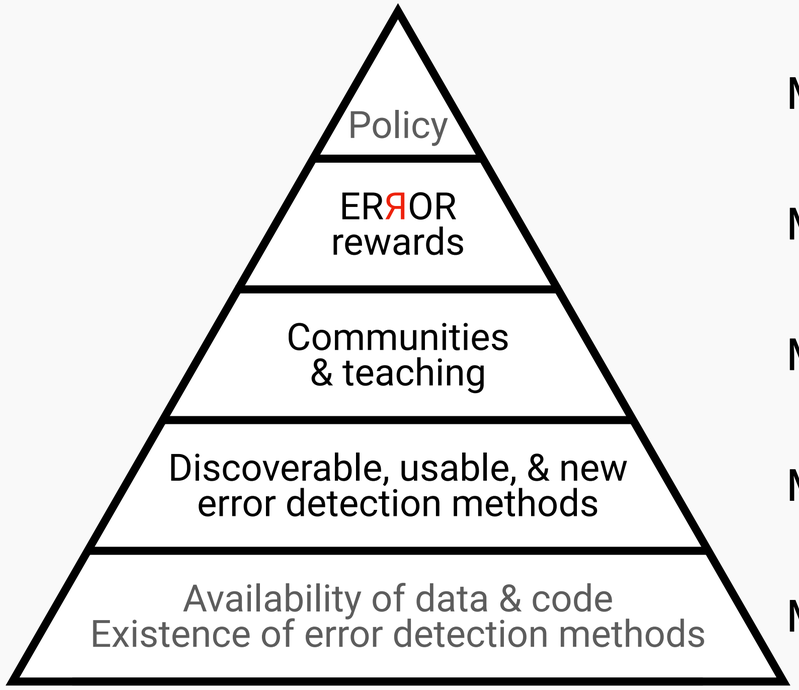
Make it normative

Make it easy

Make it possible

Nosek (2019)

Our strategy to increase error checking, reporting, & correction



Make it required

Make it rewarding

Make it normative

Make it easy

Make it possible

Make it easy

- Increasing the discoverability & usability of error detection methods

- New error detection methods

Make it normative

A master’s degree course in error detection

Make it rewarding

ERЯOR: A bug bounty program for science

UNIVERSITÄT BERN

Increasing the discoverability & usability of error detection methods

xxxxxxxxxxxxxxxxxxxxxxxxxx

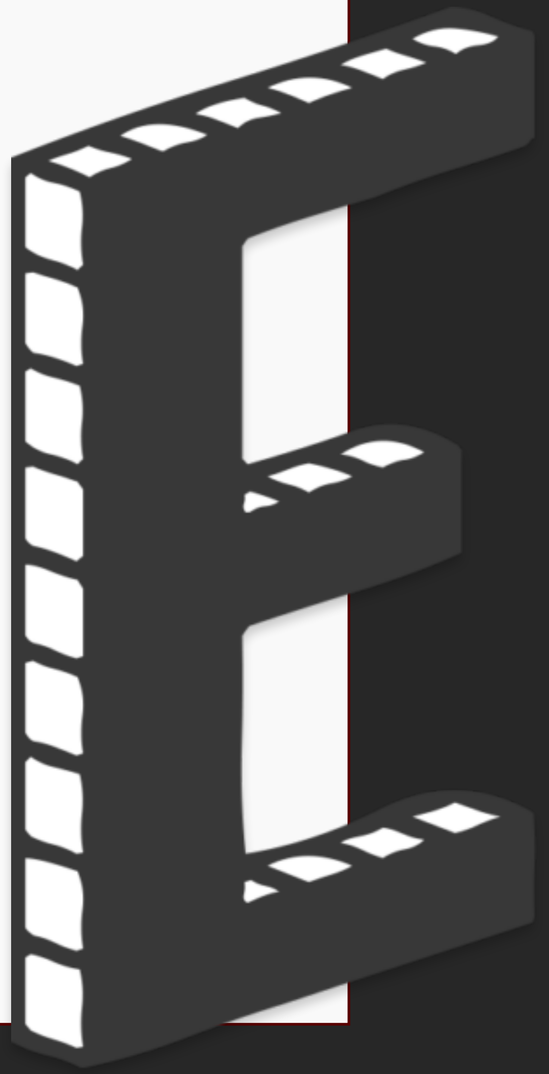
01.

01.

‘The scientific article’ is a well-defined genre

Yet “no handbooks exist for writing replies”

(Hyman, 1995)



An ecosystem of error checking methods

Methods to scrutinise the article, data, code, or the relationships among them

Smoke vs fire methods

Often

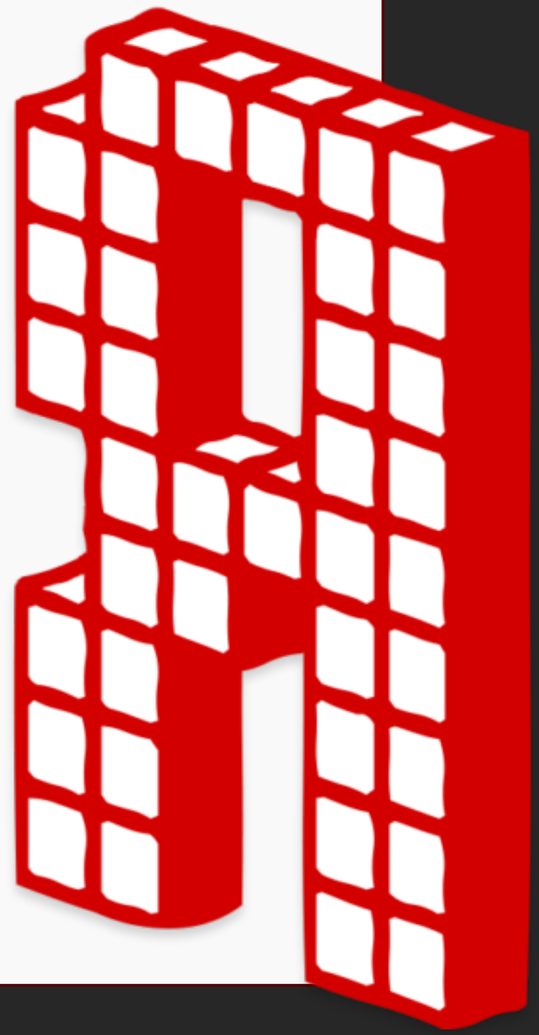
- Very simple math
- Exploiting overlooked details & repetition
- Established principles redeployed for error detection

01.

Repetition & recalculation methods

- StatCheck

“*t*(58) = 1.46, *p* = .03” → *p* = .19

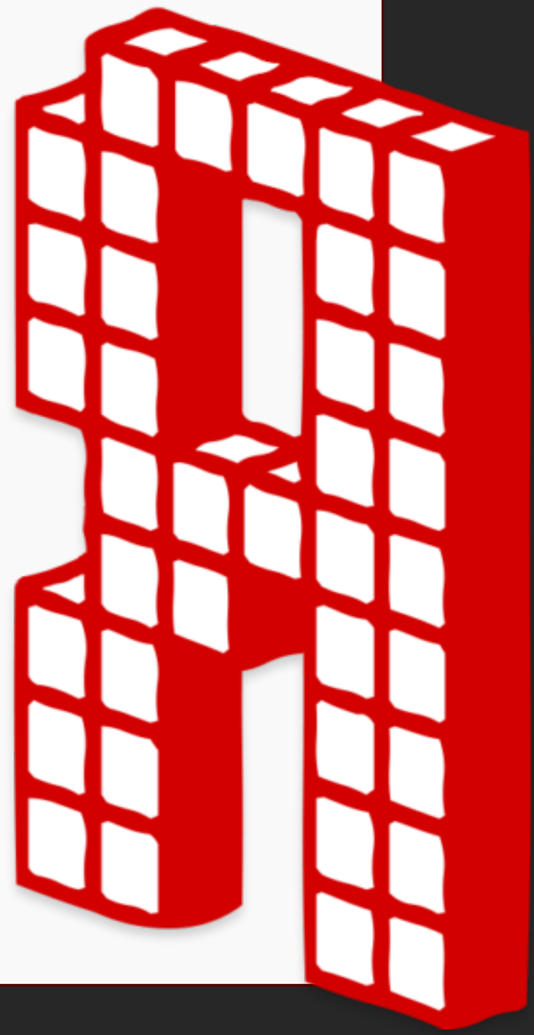


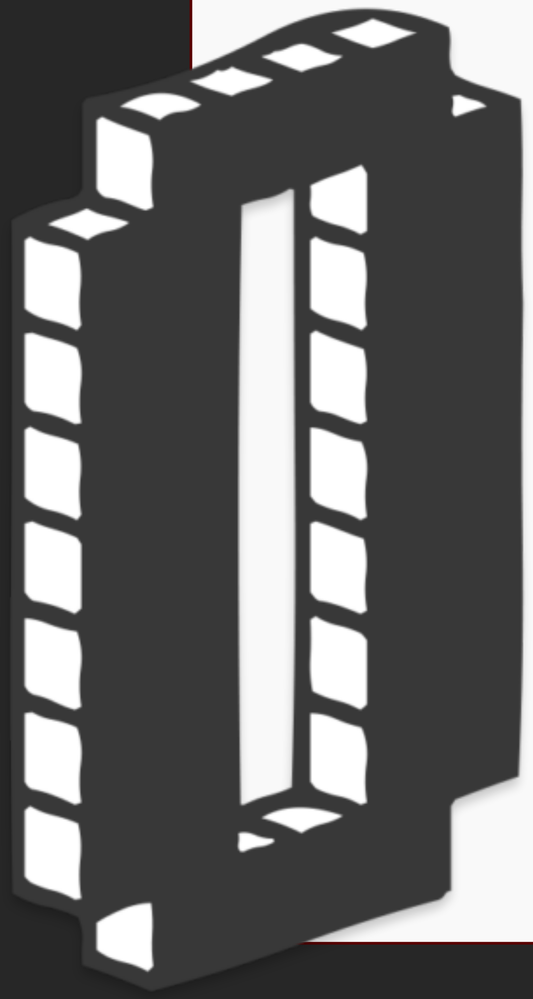
01.

Repetition & recalculation methods

- StatCheck
- GRIM / GRIMMER / GRIMMEST
- SPRITE
- RIVETS
- DEBIT
- Reproduce effect-sizes & test statistics from summary statistics
- …

R packages, vignettes, how-to guides, blogs, & examples needed!





‘Shoe-Leather’ error checks

- Full sample - exclusions = analytic sample size?
- Point estimates inside the confidence intervals?

Expert-knowledge based checks

- Beck Depression Inventory II: “SD = 3.40” is too low

Currently few training materials

03.

A foreseeable output of ERROR:

An open handbook for error detection

- What do you have?
  - Article, measures, data, code, the original authors' good-will?
- How can you get more?
  - Extracting additional data (plots, tables, summary statistics, asking)
- What could you check and how?
- How to prioritise?
  - What are the authors' claims? (Scheel, 2022)
- How to ensure reproducibility of checks?
- How to effectively communicate errors

Where do you want to look?

In the paper only

What resources do you have available?

- [x] The paper
- [ ] The code
- [ ] The raw data
- [ ] The first author's home address

What do you care about?

- [x] Statistical errors
- [ ] Conceptual errors
- [x] Reporting errors
- [ ] Integrity errors
- [ ] Failing-to-cite-me errors

How much time do you have?

< 2 hours

Error-checking tools for you:

- GRIM
- GRIMMER
- SPRITE
- TIDES
- PORT
- STALT

drjamiecummins.com/error-selector-demo

UNIVERSITÄT BERN

New error detection methods

xxxxxxxxxxxxxxxxxxxxxxxxx

02.

02.

What's wrong with these summary stats?

(Can't use GRIM/MER because N is too high)

"145 participants responded using a 5-item Likert scale (response options 1-7, sum scored), M = 38.1, SD = 4.2)"

"200 participants responded using a single-item Likert scale (response options 1-7), M = 2.1, SD = 2.4"

02.

What's wrong with these summary stats?

(Can't use GRIM/MER because N is too high)

"145 participants responded using a 5-item Likert scale (response options 1-7, sum scored), M = 38.1, SD = 4.2)"

➔ Mean must be within [5, 35]

"200 participants responded using a single-item Likert scale (response options 1-7), M = 2.1, SD = 2.4"

➔ SD must be within [0.31, 2.31]

Mestdagh et al. (2018)
Hussey et al. (in prep)

TIDES | Analyze a single set of summary stats | Upload and analyze mutiple sets | Explanation & Links

mmmdata.shinyapps.io/TIDES

Truncation-Induced Dependency in Summary Statistics

A method for checking the compatibility of reported means, SDs, and Ns given the min and max of the scale

Mean:

2.1

Standard Deviation:

2.4

Sample Size:

30

Minimum Value:

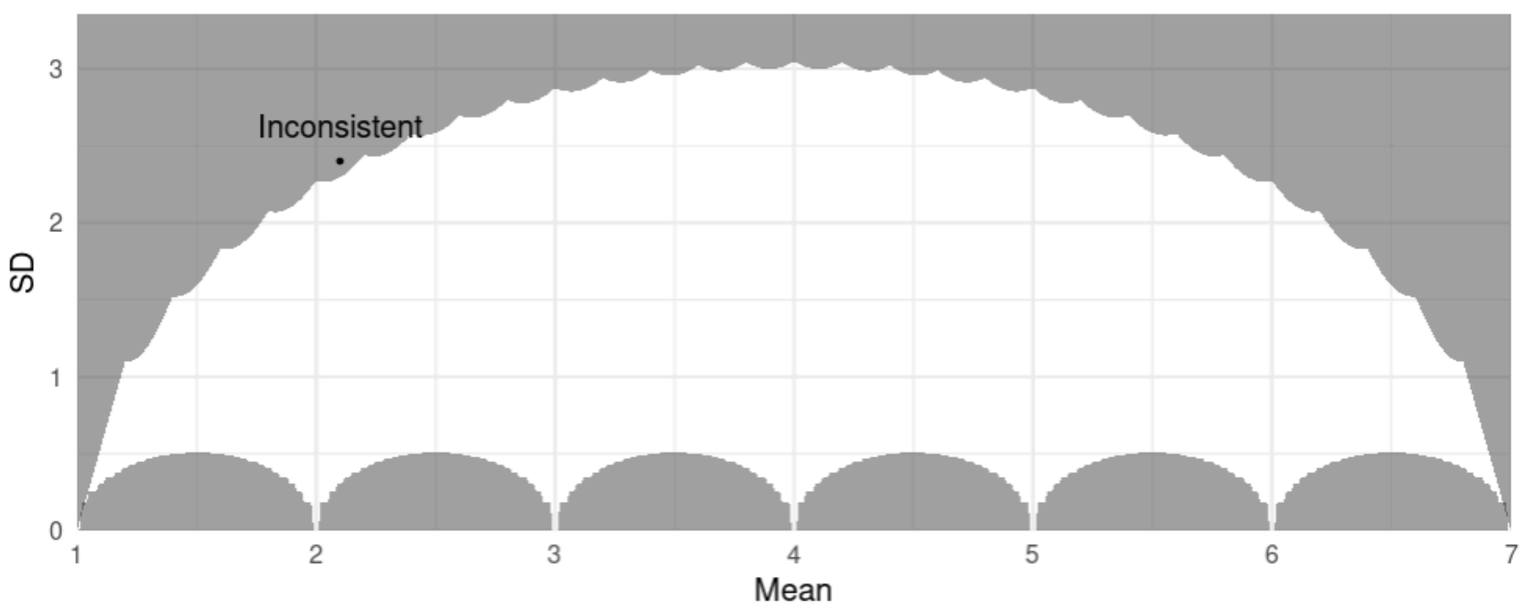
1

Maximum Value:

7

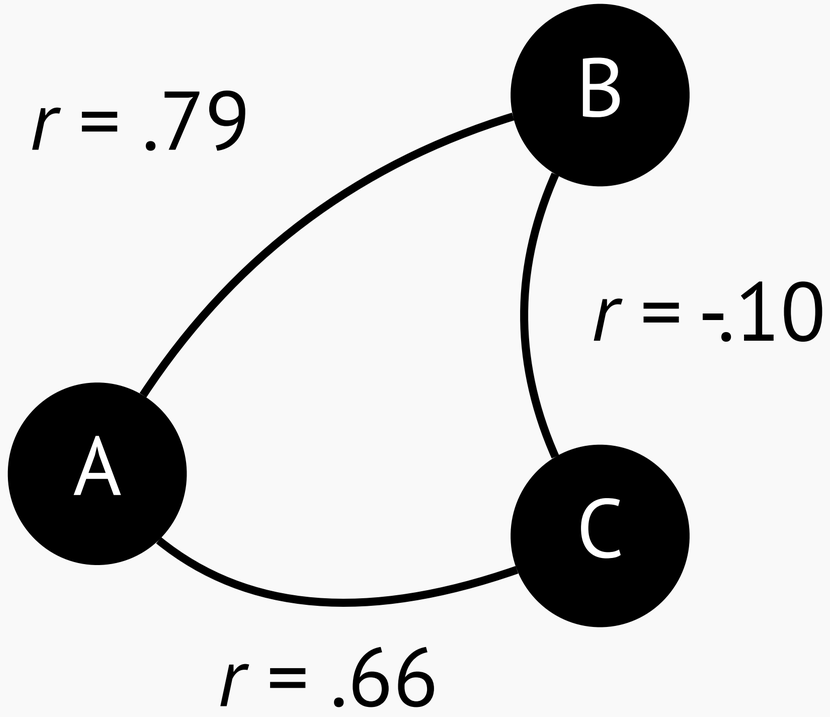
Scale is interval:

☑ Yes, also calculate minimum SD

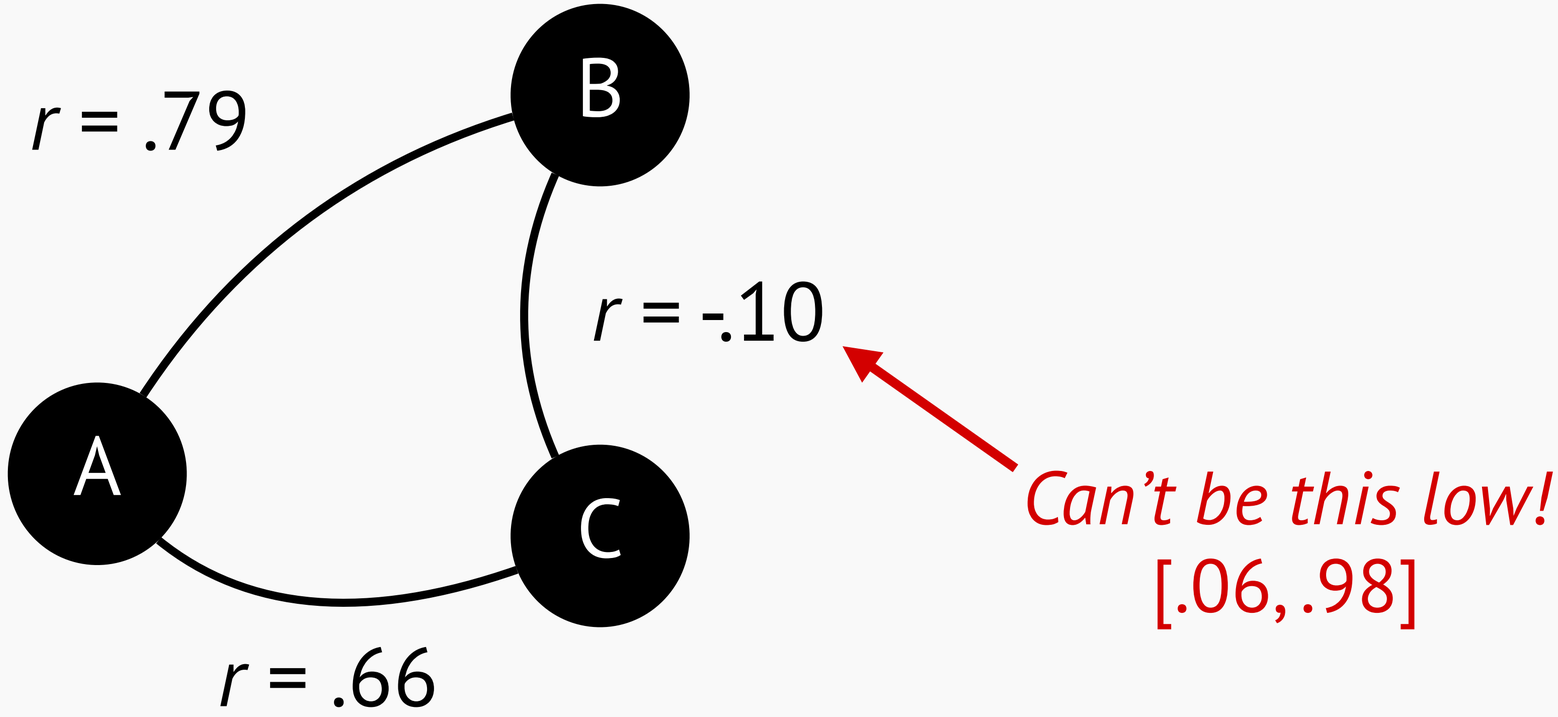


02.

What's wrong with these correlations?



What's wrong with these correlations?



Stanley & Wang (1969)

02.

Does this correlation table contain errors?

Hard to spot by eye; easier with tools

X1	X2	X3	X4	X5	X6	X7	X8
	0.39	0.80	0.70	-0.14	0.19	-0.13	0.46
		0.43	0.10	-0.03	0.12	-0.14	0.72
			0.08	-0.13	0.17	-0.11	0.41
				0.01	0.17	-0.09	0.08
					0.06	0.02	-0.03
						-0.11	0.14
							-0.13

PORT | Analyze a set of three correlations | Upload and analyze a full correlation matrix | Explanation & Links

mmmdata.shinyapps.io/PORT

Positive-definiteness Of (Pearson's) *r* Tables

A method for assessing inconsistencies among reported correlations

Download Example CSV

Choose CSV or Excel File

Browse... example_correlation_matrix.csv

Upload complete

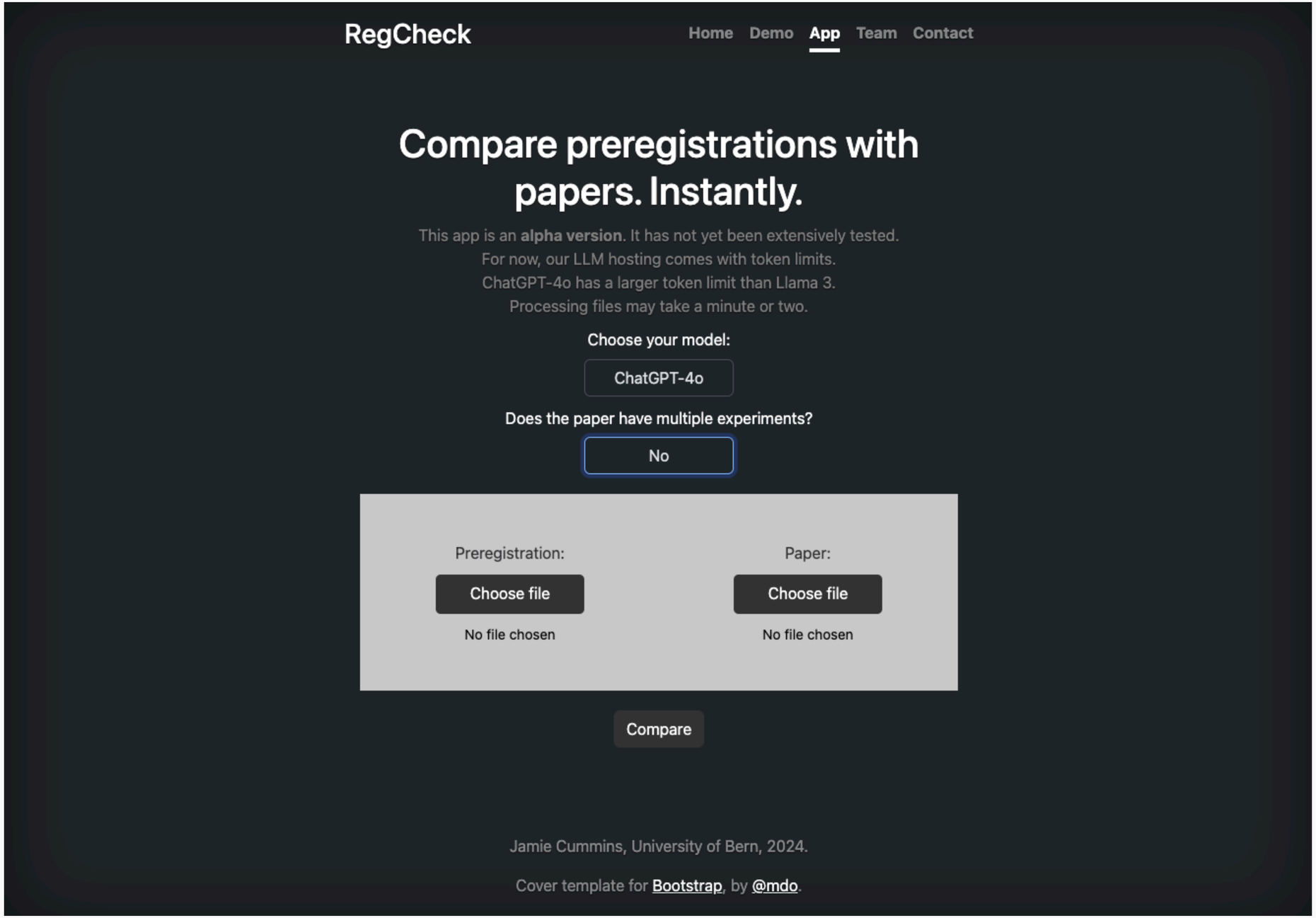
Download Results

Test Results | Correlation Imputation

This table displays the reported correlations. The difference between each reported correlation and its imputed value is shown in square brackets. E.g., if imputation suggests that a given correlation may be 0.30 larger than the reported value, this is displayed as [+0.30]. The three correlations with the largest absolute differences are annotated ***, **, and * in decending order.

V1	V2	V3	V4	V5	V6	V7	V8
NA	0.39 [+0.06]	0.80 [-0.55]**	0.70 [-0.62]***	-0.14 [+0.05]	0.19 [+0.04]	-0.13 [-0.01]	0.46 [-0.10]
NA	NA	0.43 [-0.08]	0.10 [+0.03]	-0.03 [-0.03]	0.12 [-0.03]	-0.14 [+0.04]	0.72 [-0.52]*
NA	NA	NA	0.08 [+0.46]	-0.13 [-0.07]	0.17 [-0.05]	-0.11 [+0.01]	0.41 [+0.15]
NA	NA	NA	NA	0.01 [-0.09]	0.17 [-0.06]	-0.09 [+0.01]	0.08 [+0.17]
NA	NA	NA	NA	NA	0.06 [-0.09]	0.02 [-0.00]	-0.03 [-0.02]
NA	NA	NA	NA	NA	NA	-0.11 [+0.08]	0.14 [-0.02]
NA	NA	NA	NA	NA	NA	NA	-0.13 [+0.01]
NA	NA	NA	NA	NA	NA	NA	NA

regcheck.app



UNIVERSITÄT BERN

A master’s degree course in error detection

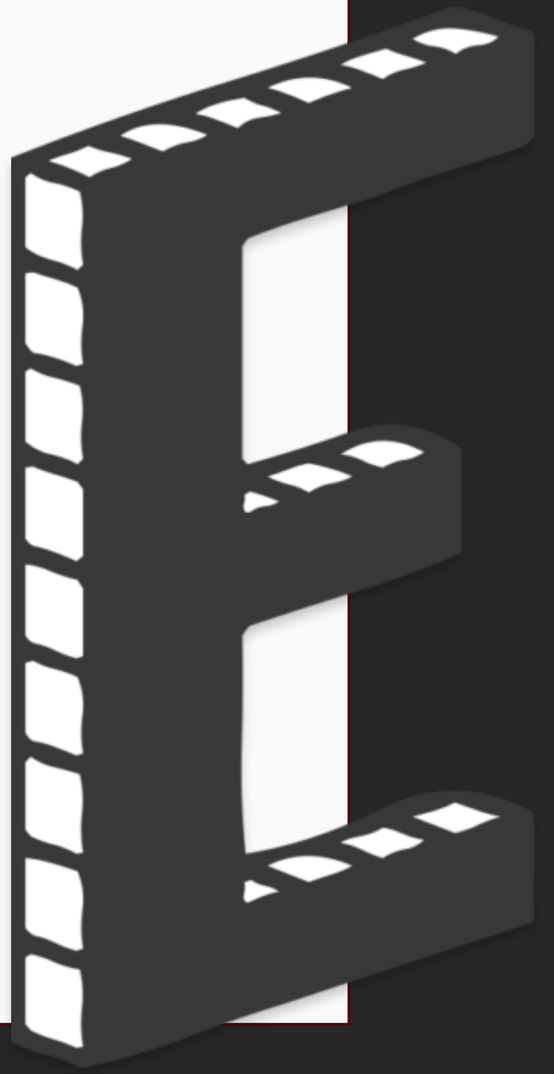
xxxxxxxxxxxxxxxxxxxxxxxxx

03.

03.

Conspicuous absence of error detection in the curriculum

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



03.

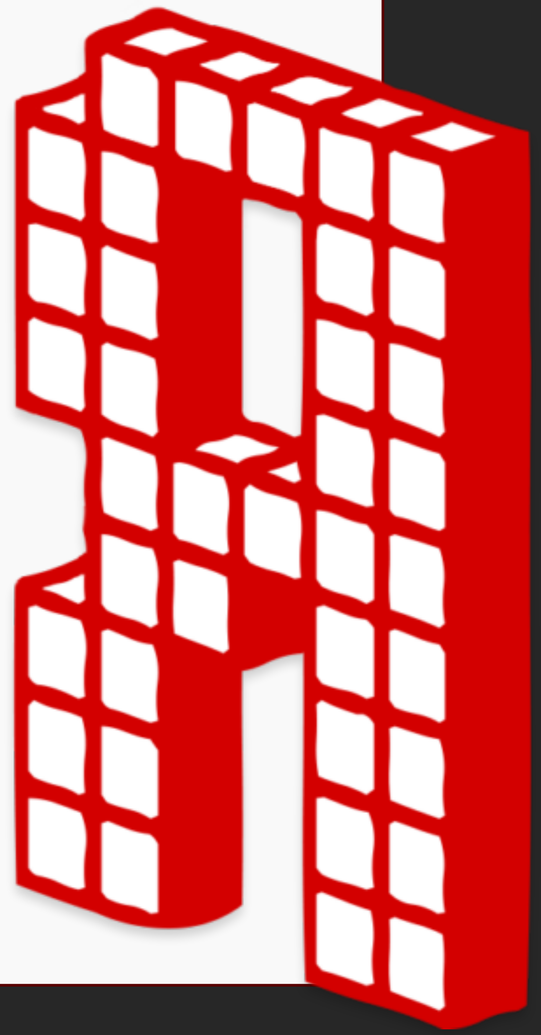
Masters course:

“Estimating the credibility of past research”

(@UniBe since Fall ’23)

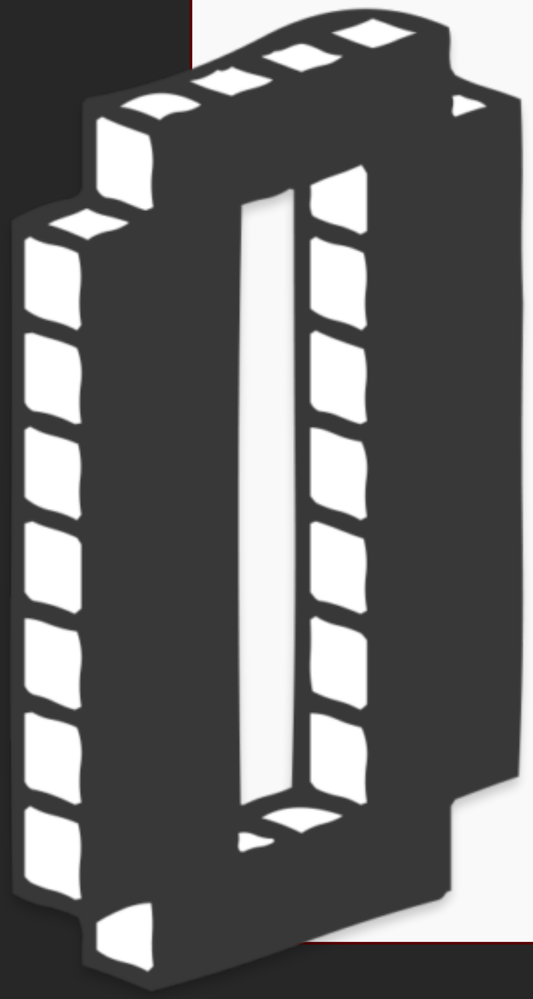
Science as an intensely human & fallible activity

Organised skepticism



Goals

- Explore and test **practical challenges** in the implementation of an error checking system
- Estimating a **benefit-cost ratio** of an error detection system relative to **not detecting** these errors
- Obtaining empirical estimates of the **types** of errors & their **prevalence**

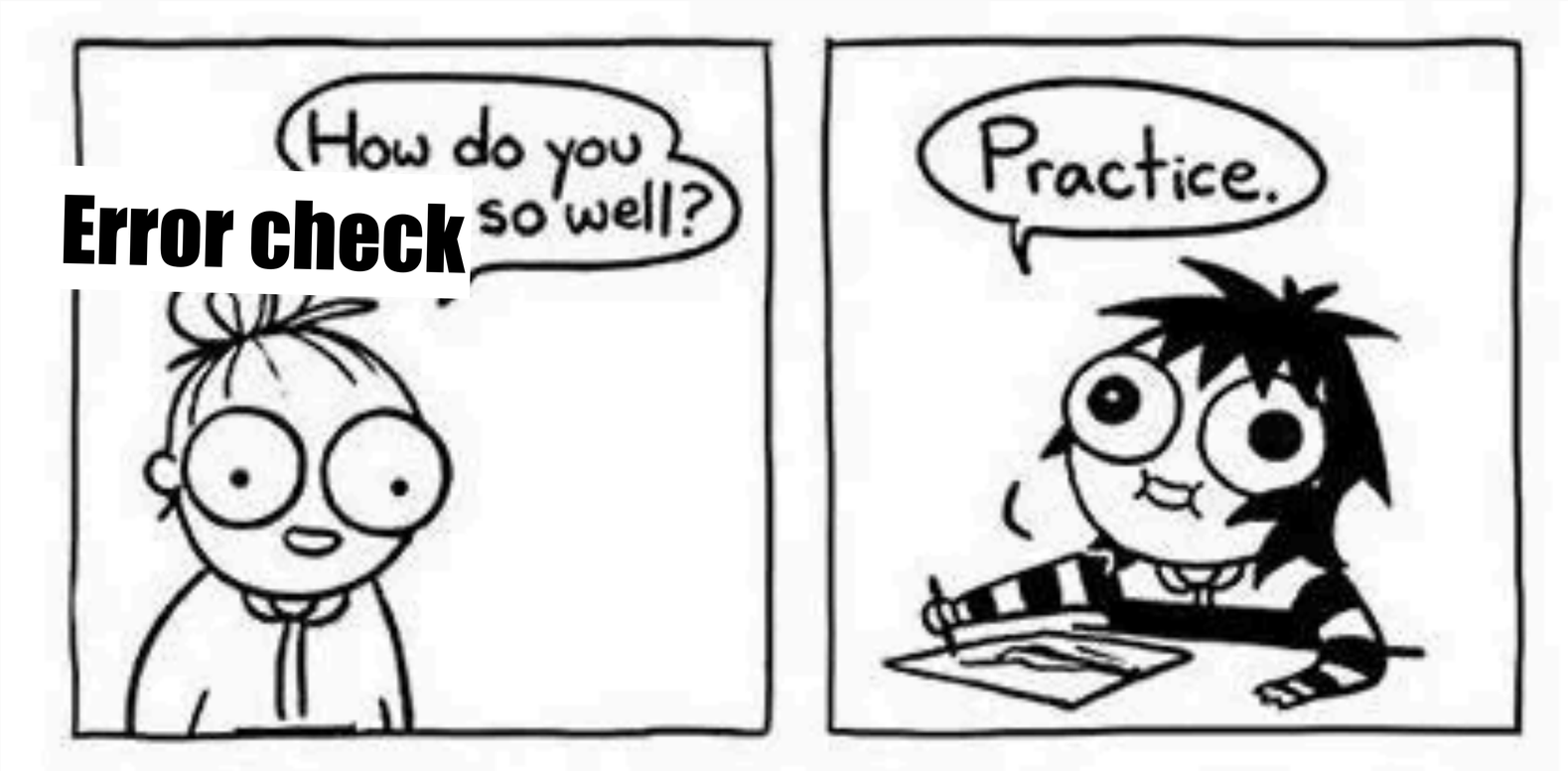


Syllabus

- Mentality: Functions of science other than truth discovery
- *p*-curve, StatCheck, GRIM, GRIMMER, SPRITE, TIDES, PORT, …
- Recomputing results from summary statistics
- Effect size plausibility
  - Reliability corrections
- Meta-analysis bias detection & its failures
- (Missing) causal language
- Spotting jingle/jangle fallacies
- Obtaining & extracting data
- Communicating errors + examples of critique

Not merely information, practice!

03.



Error check

‘Normalise’ is a verb

03.

Problems encountered

Permission to critique

xxxxxxxxxx

Language of critique

Exposure & Practice!

UNIVERSITÄT BERN

ERЯOR

A bug bounty program for science

xxxxxxxxxxxxxxxxxxxxxxxxxx

04.

£220,000

(250k CHF, €260k, $285k)

fund

Check

100

published articles
for errors

Pay authors & reviewers

+ bonus reward

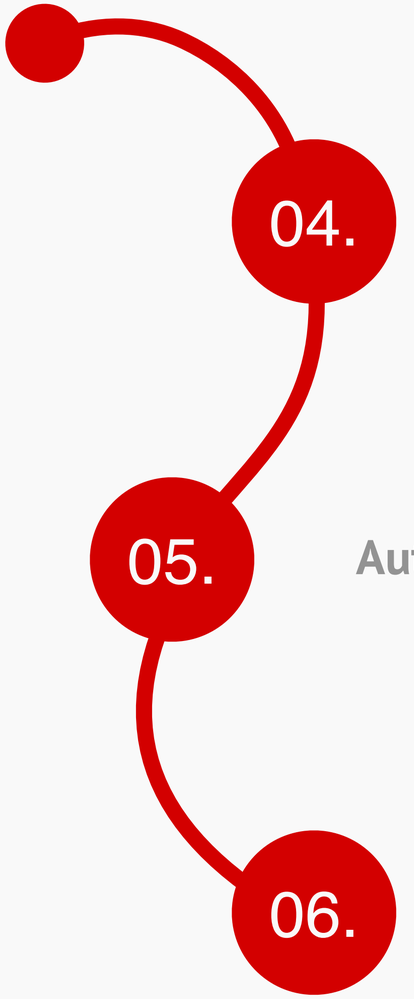
contingent on errors found

Candidate articles

Obtain author consent

Match with reviewer

- Selected for ‘importance’:
  - Would matter if it contained errors
  - 30+ citations per year
- Recent (<10 years)
- Attention paid to representation & power
  - Gender, precarious employment, etc.



Error report

Author response

Decision, recommendation, & reward payments

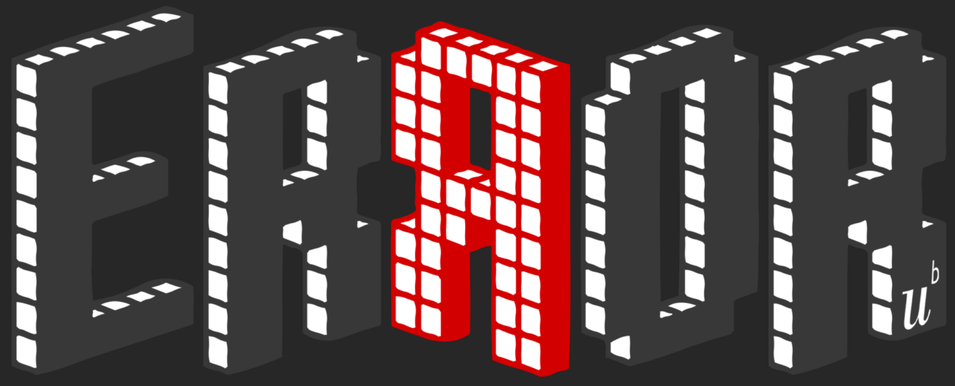
- Recommender makes decision & recommendation
  - All published on error.reviews website
  - CC-By 4.0 licence
- Possibly in conjunction with the advisory board

	Base	**Bonus reward**
Author	£ 220	Up to £220
Reviewer	£220 – £1,100 Depending on expected effort	Up to £2,200

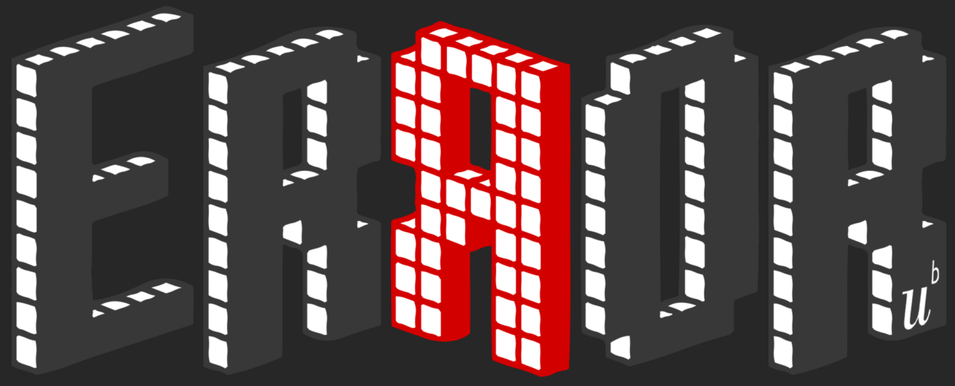
* *Indicative ranges; will be scaled for regional differences*

Bonus rewards

Decision	**Recommendation**	**Bonus reward** (indicative)
No errors	No additional action beyond publication of error report on the error.reviews website.	£220 to the author
Minor errors	Authors to appropriately recognise these errors in future discussions of the article. *Probably most research!*	£220 to both the author & reviewer
Indeterminable errors	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.	£220 to the reviewer
Moderate errors	Correction notice (minor)	£440 to the reviewer
Major errors	Correction notice (major) / may warrant an expression of concern	£880 to the reviewer
Severe errors	Retraction	£2,200 to the reviewer

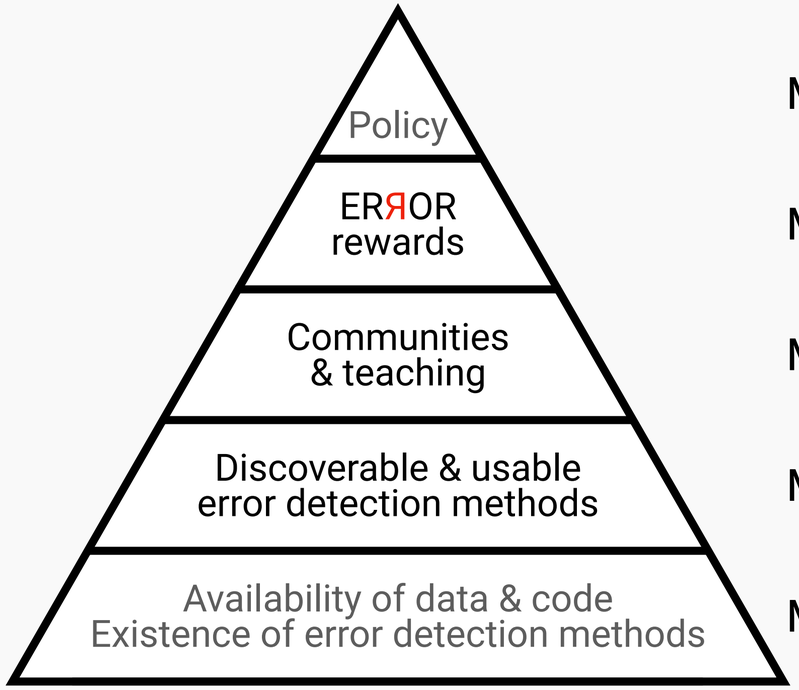


1 / 14	134	1	1'250
			/ 250'000 CHF
Reviews Completed / Pending	Invited Papers	Papers with Errors	Total Payouts



error.reviews

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
Twitter: @ianhussey
Bluesky: @ianhussey.bsky.social

error-reviews.psy@unibe.ch
Twitter: @error_reviews
Bluesky: @error.reviews

UNIVERSITÄT BERN

xxxxxxxxxxxxxxxxx

Collaborators

- Malte Elson
- Ruben Arslan
- Jamie Cummins
- Sabrina Norwood
- Frank Bosco
- Imran Kadolkar

Links

- ERROR reviews error.reviews
- PORT mmmdata.shinyapps.io/PORT
- TIDES mmmdata.shinyapps.io/TIDES
- Finder drjamiecummins.com/error-selector-demo
- RegCheck regcheck.app

xxxxxxxxxxxxxxxxx

Other normalisation efforts:

“Commentaries don’t get cited”

	Citations of target article	Citations of commentary	Within-pair proportion
Mean	106	5	25%
Median	80	10	15%
Minimum	2	1	1%
Maximum	490	53	150%

N = 36 commentaries published in Psychological Science between 2007-2021