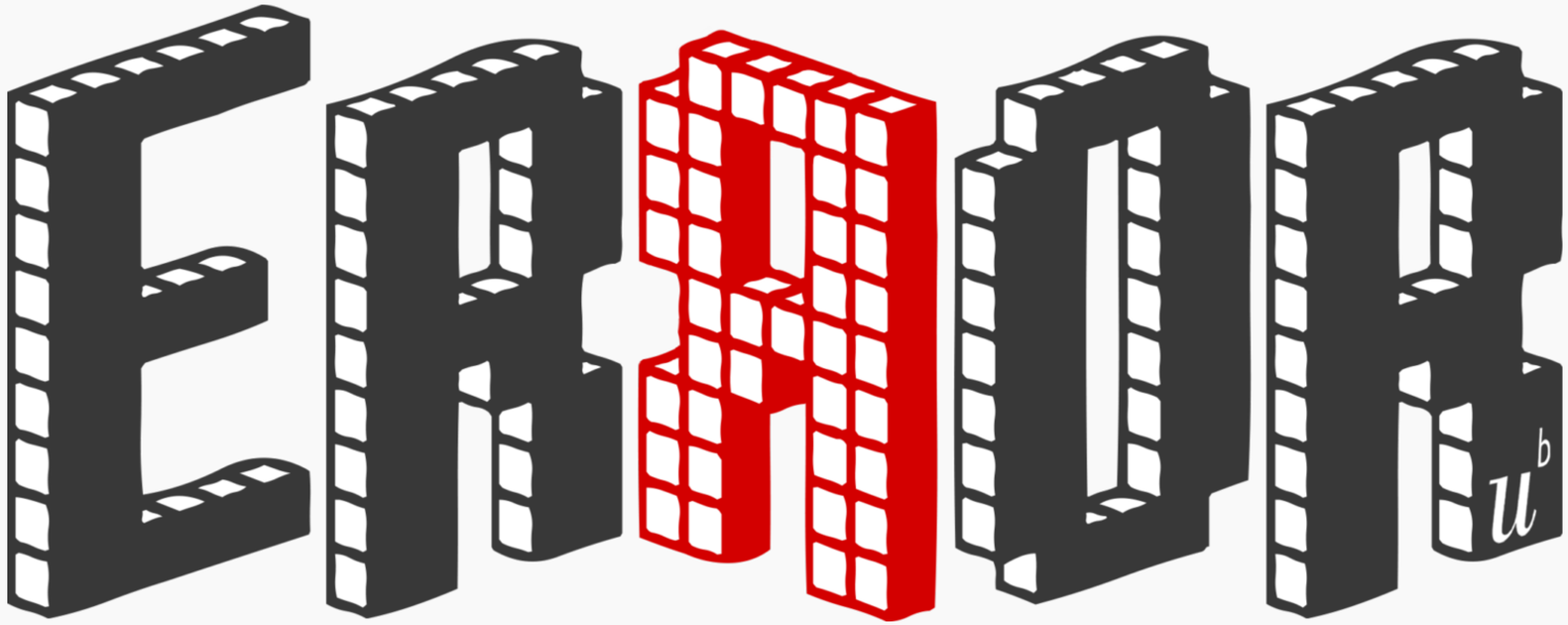
UNIVERSITÄT BERN

xxxxxxxxxxxxxxxxxxxxxxxxxxxx



xxxxxxxxxxxxxxxxxxxxxxxxxxxx

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

Ian Hussey

UNIVERSITÄT BERN

PRINCIPAL INVESTIGATORS

Malte Elson
University of Bern

Ruben Arslan
University of Leipzig

CHIEF RECOMMENDER

Ian Hussey
University of Bern

UNIVERSITÄT BERN

ADVI2ORY BOARD

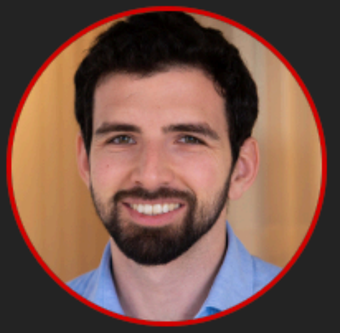
Dorothy Bishop
University of Oxford

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

Errors are broadly conceived

‘An act that through ignorance, deficiency, or accident departs from or fails to achieve what should be done’

Errors are broadly conceived

- Includes both provable & probable errors
- Goes beyond computational reproducibility
- Hard to check errors and not also scrutinize more generally, eg flexibility, robustness, measurement

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

Who can help carry this burden?

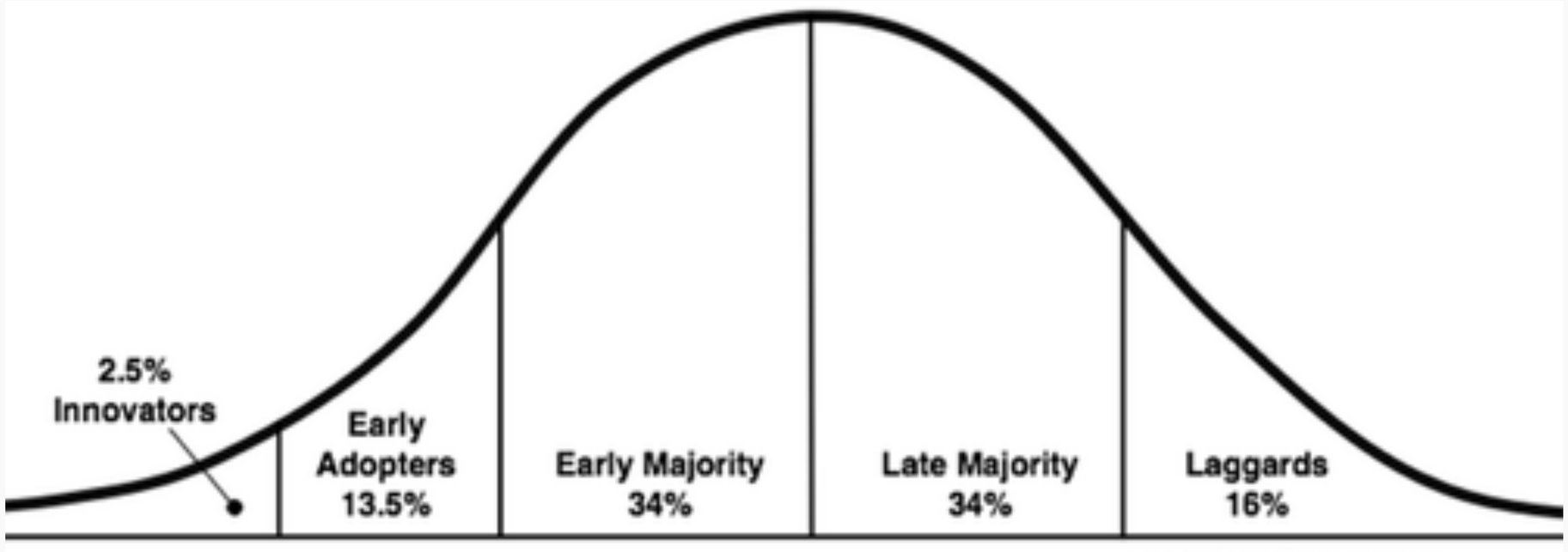
"A few outsiders and weirdos saw the giant lie at the heart of [science], and they saw it by doing something the rest never thought to do:

They looked."

Lessons from the Open Science/methods reform movement

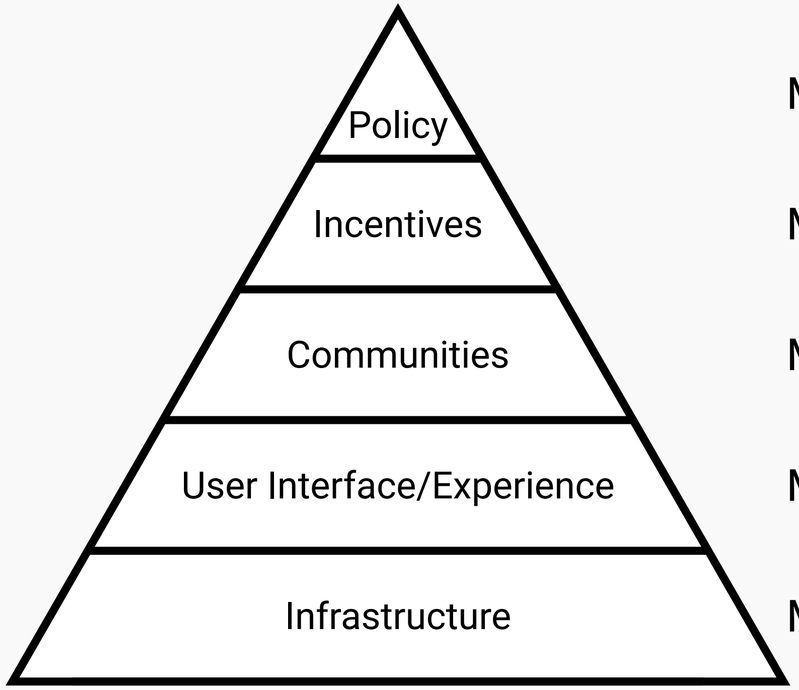
- Thought leaders and experts are important
- But reliance on a few diligent experts isn't scalable
- You also need a grassroots, ECR-driven movement
- And probably a stack of money too

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



Nosek (2019)

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



Make it required

Make it rewarding

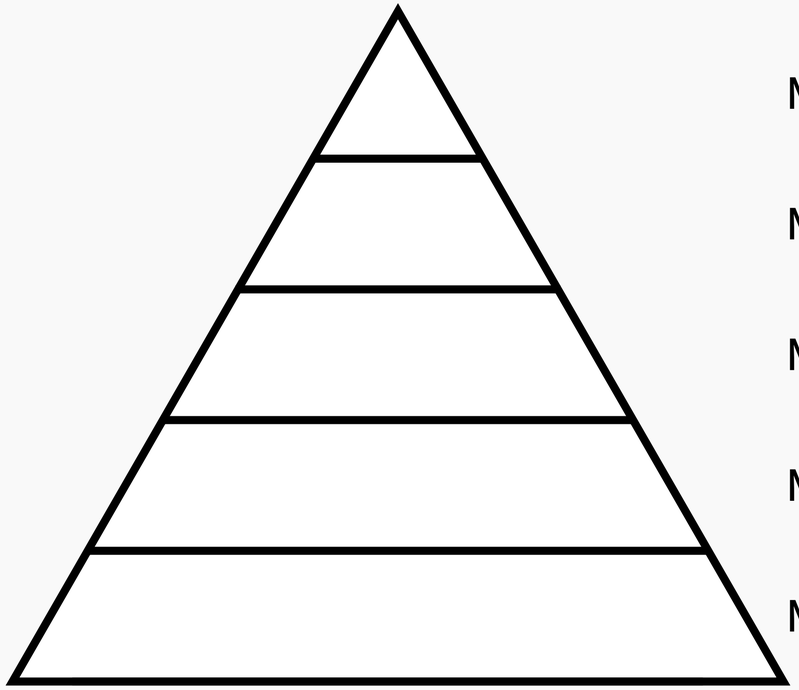
Make it normative

Make it easy

Make it possible

Nosek (2019)

Our strategy to increase error checking, reporting, & correction



Make it required

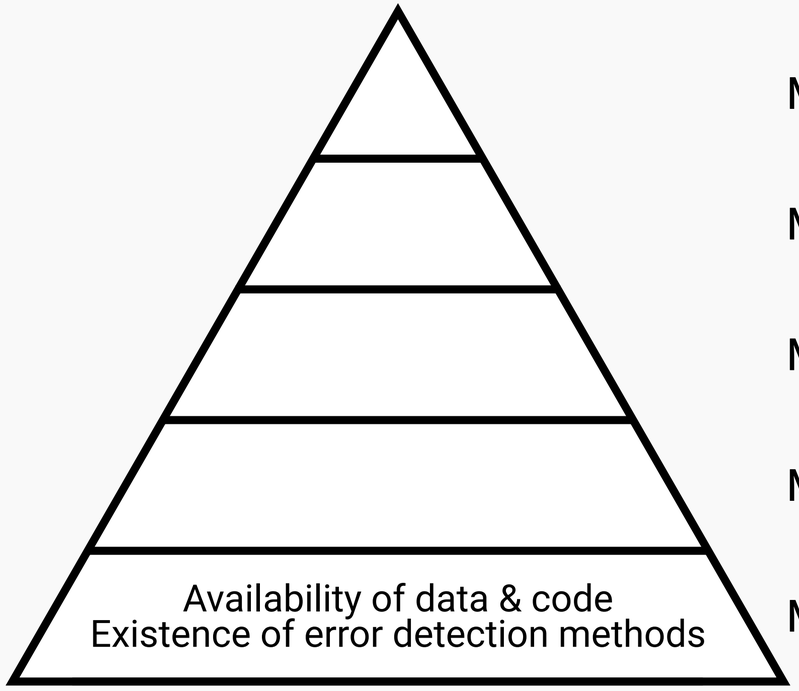
Make it rewarding

Make it normative

Make it easy

Make it possible

Our strategy to increase error checking, reporting, & correction



Make it required

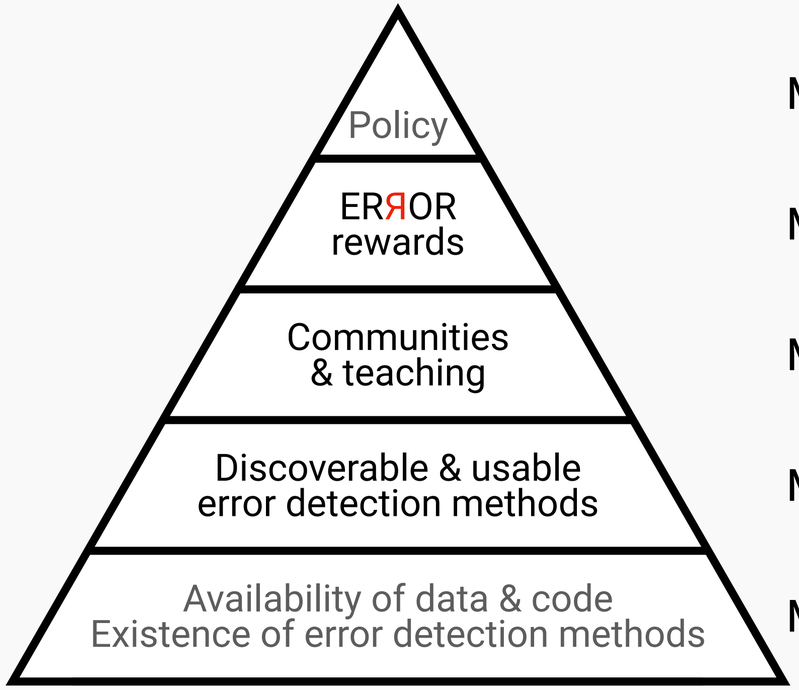
Make it rewarding

Make it normative

Make it easy

Make it possible

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

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.

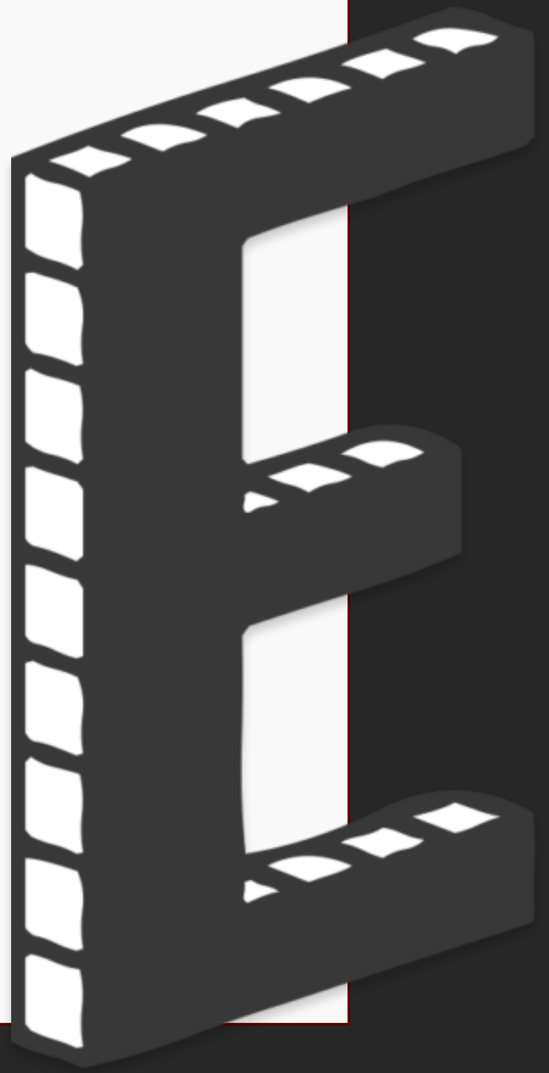
01.

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



‘The scientific article’ is a well-defined genre

“No handbooks exist for writing replies”
(Hyman, 1995)

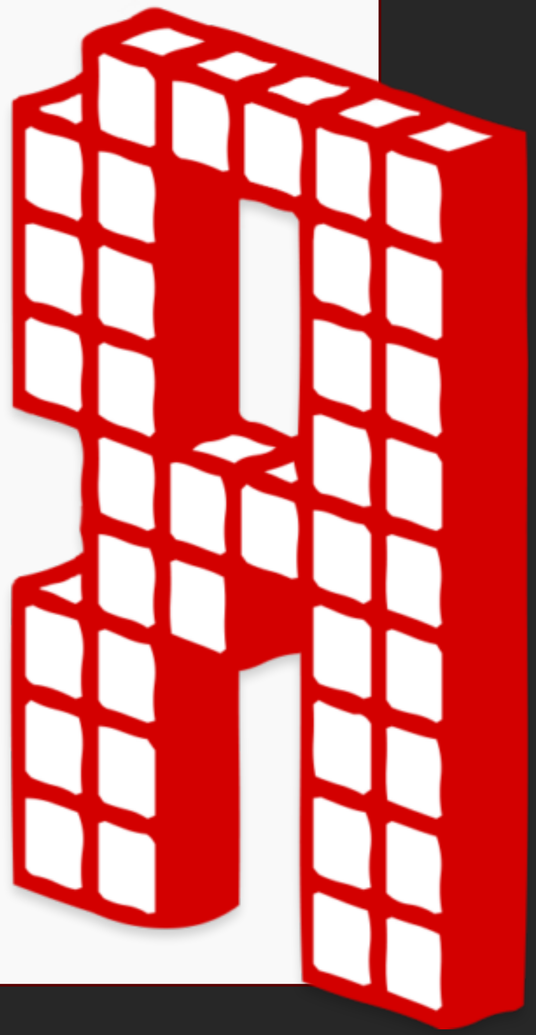
…maybe we need to write one

01.

Repetition & recalculation methods

- StatCheck

“$t(58) = 1.46, p = .03$”

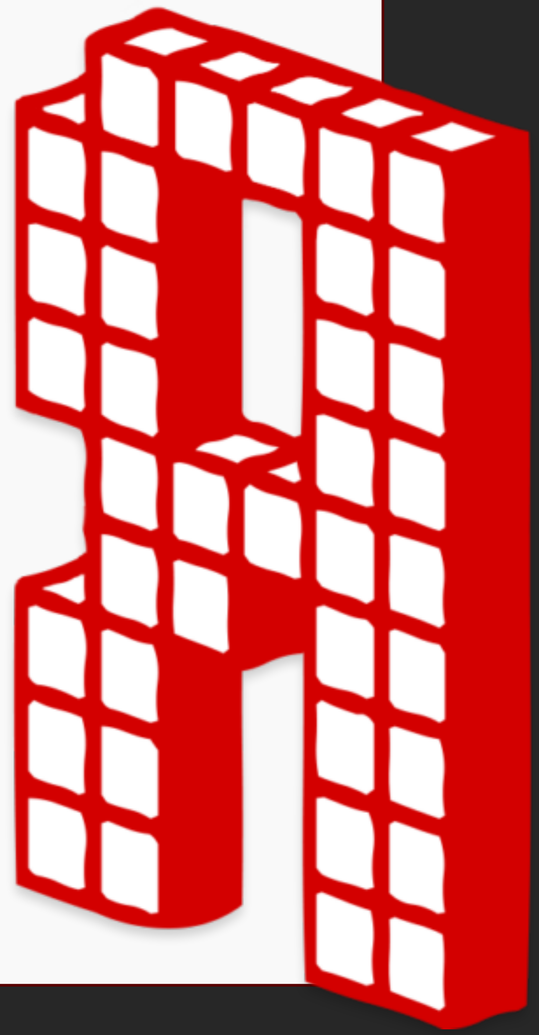


01.

Repetition & recalculation methods

- StatCheck

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



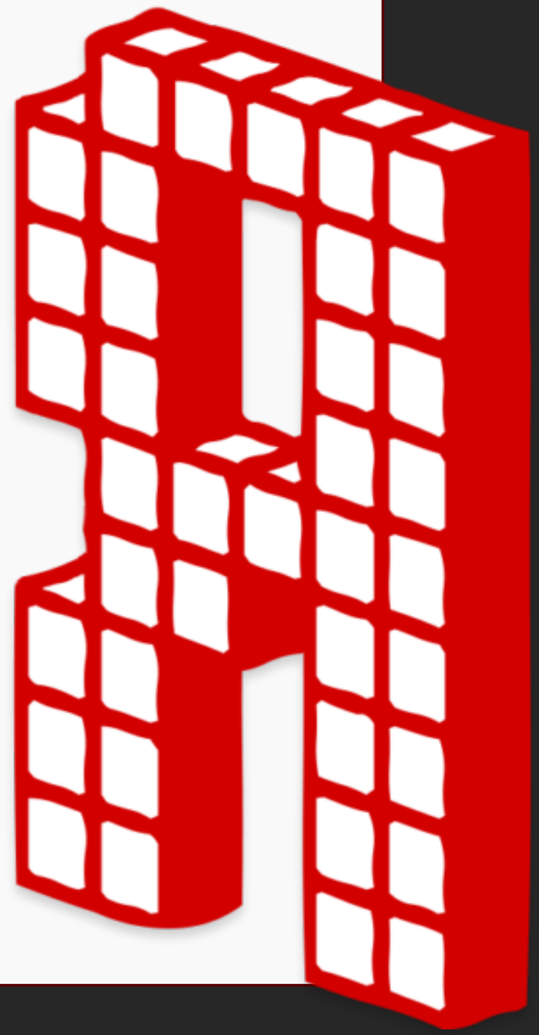
01.

Repetition & recalculation methods

- StatCheck
- GRIM

“two people rated the statement’s believability on a 1-5 scale (mean = 2.7)”

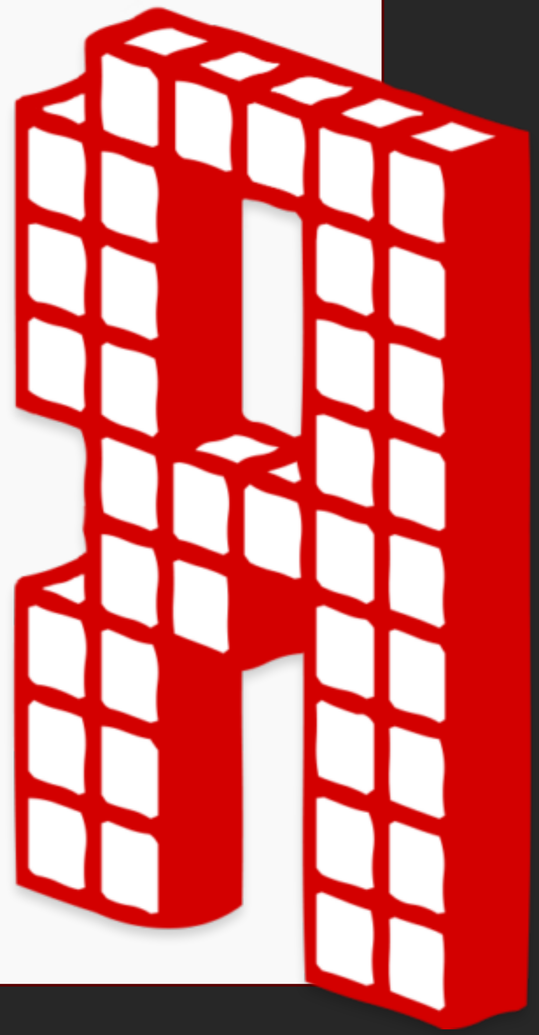
➔ *mean must end in .0 or .5*



01.

Repetition & recalculation methods

- StatCheck
- GRIM / GRIMMER / GRIMMEST
- SPRITE

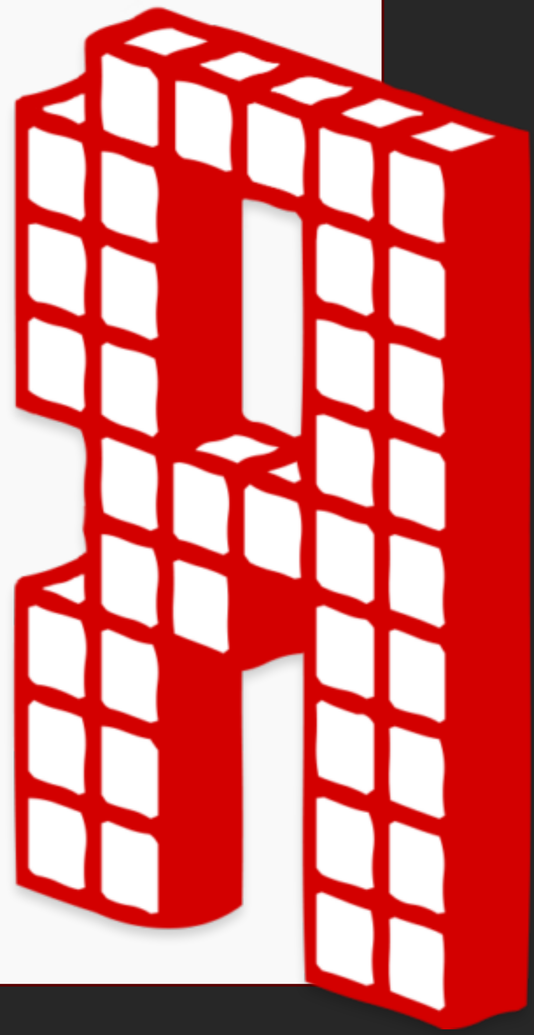


01.

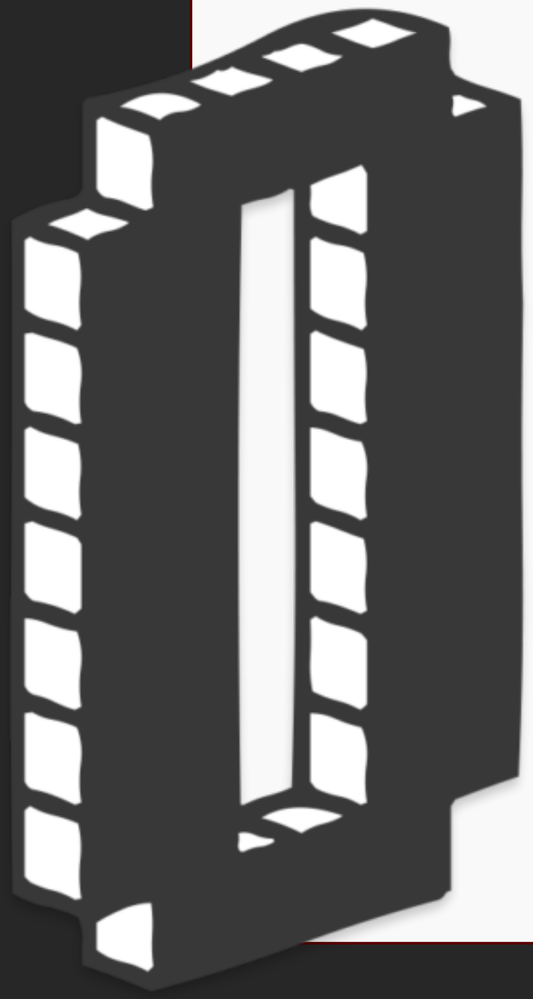
Repetition & recalculation methods

- 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 checking

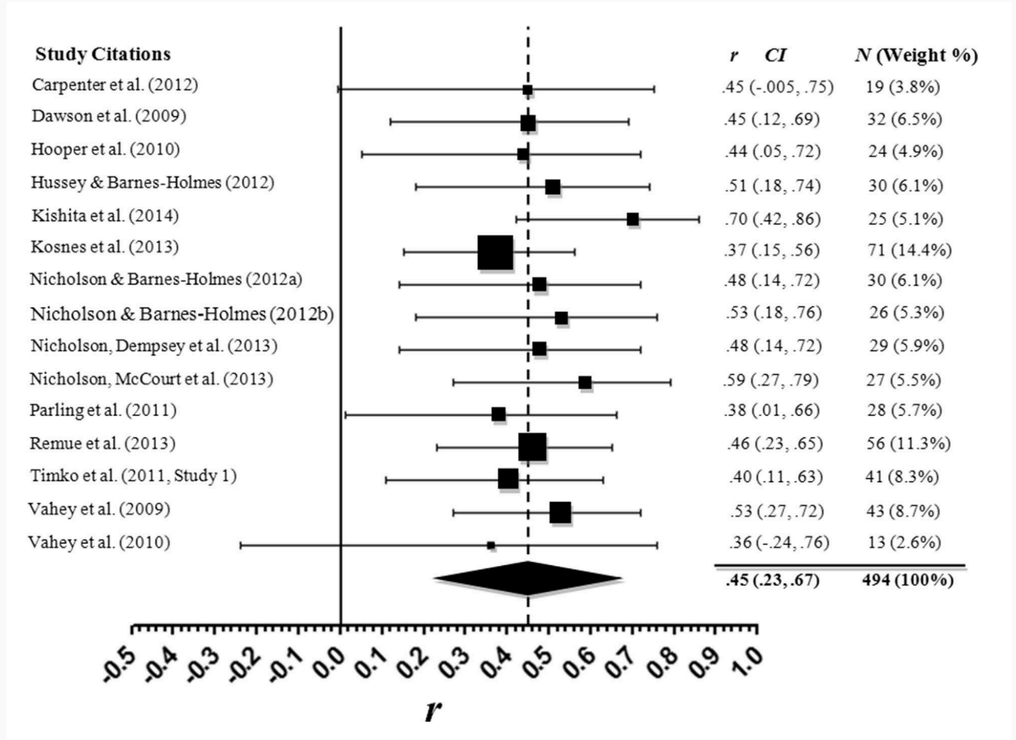


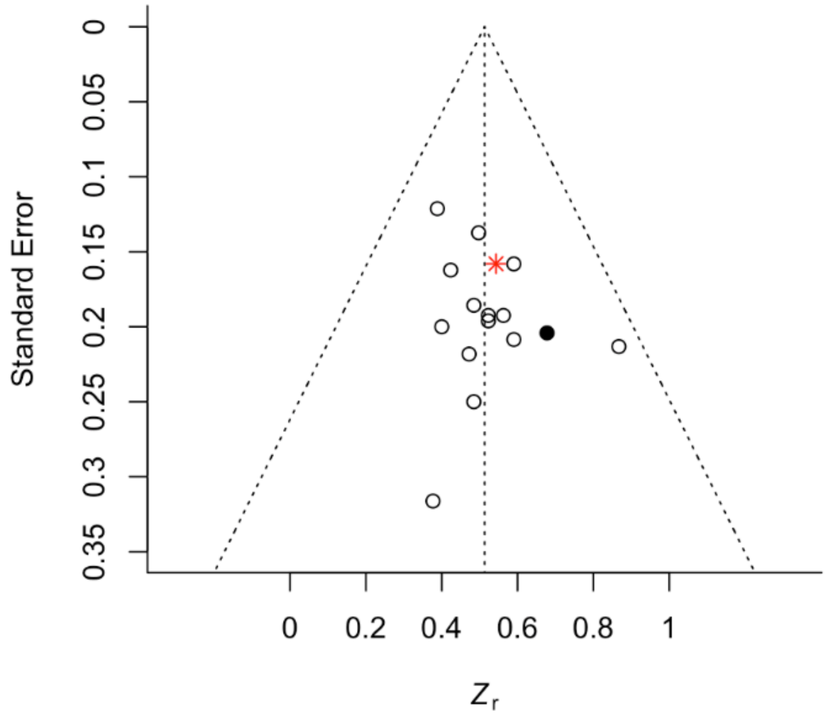
Unfortunate lack of acronyms!

- Full sample - exclusions = analytic sample size?
- Means inside scale range?
  - “Sum scores on the five-item scale (response options 1-7) … Mean = 38.1” ➔ [5, 35]
- Point estimates inside the confidence intervals?
  - “β = .11, 95% CI [.27, .53]”
- …

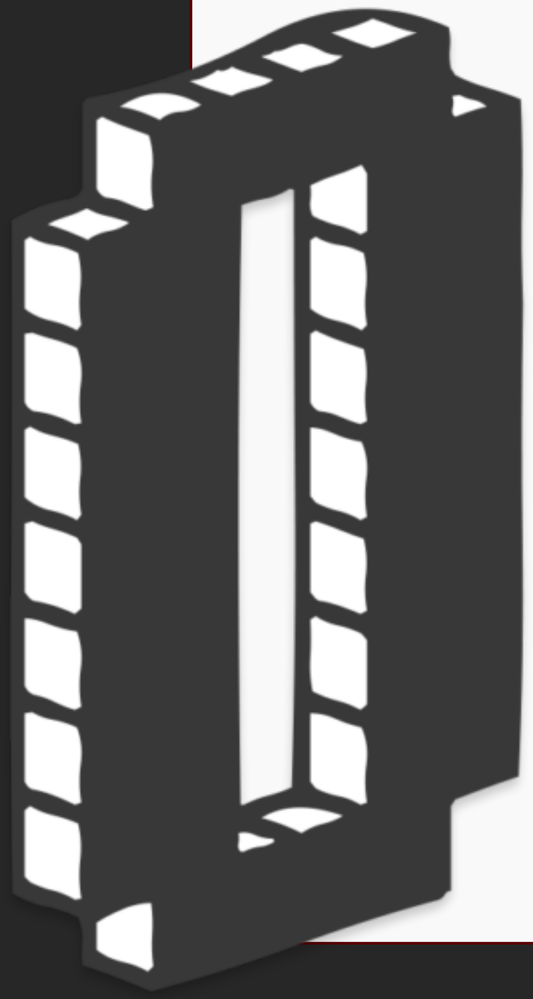
Nick Brown is king of this - but there are no training materials

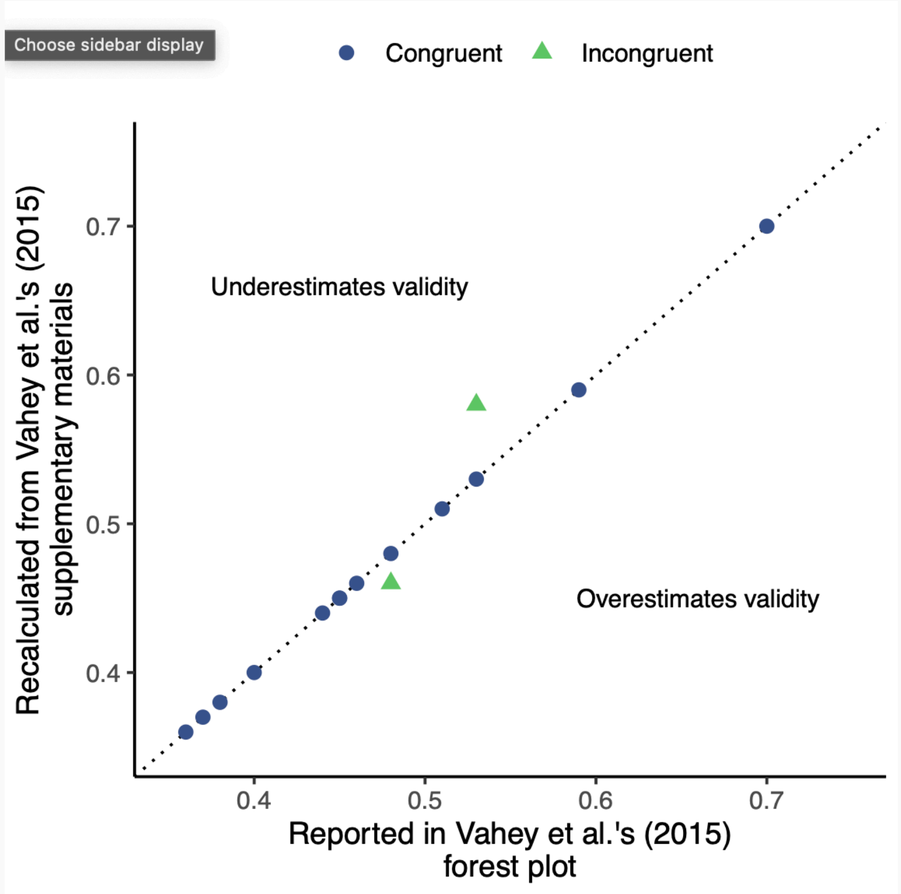
- Forest plot = funnel plot?
  - From Vahey et al. (2015)





- Meta-analysis effect sizes = effect sizes in original paper?
  - See Maassen et al. (2020)





01.

Inferential & argumentation errors

- Citation errors
  - 9% completely mischaracterised (Cobb et al, 2023)
- Difference between signif. & non-signif. is not itself significant (Gelman, 2006)
- Reverse causality
  - Suicide literature: current depression “predicts” historic suicide attempt
  - Confusing DV/IV (Fried & Kievit, 2015)
- …

01.

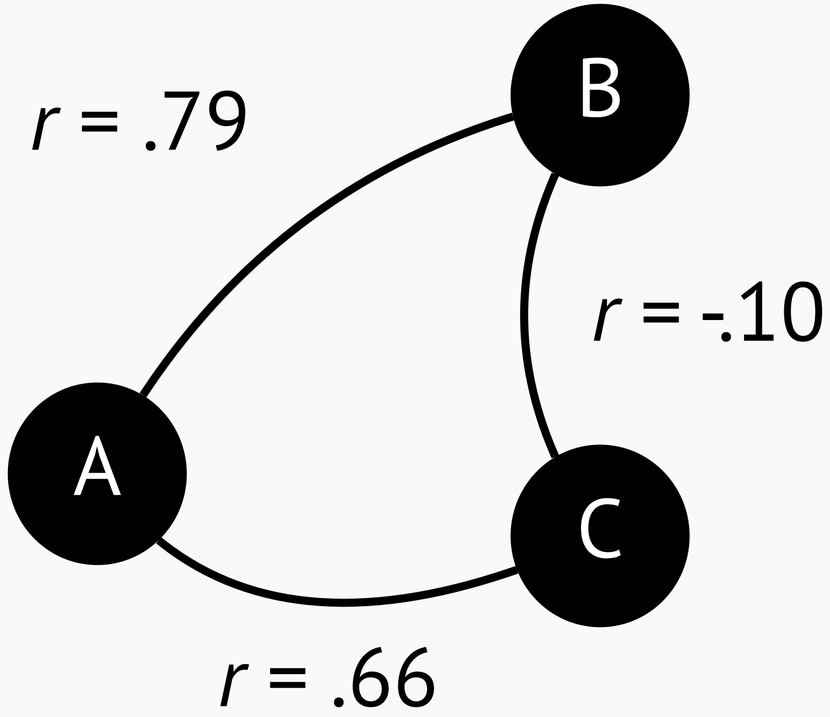
A foreseeable output:

A handbook for error detection

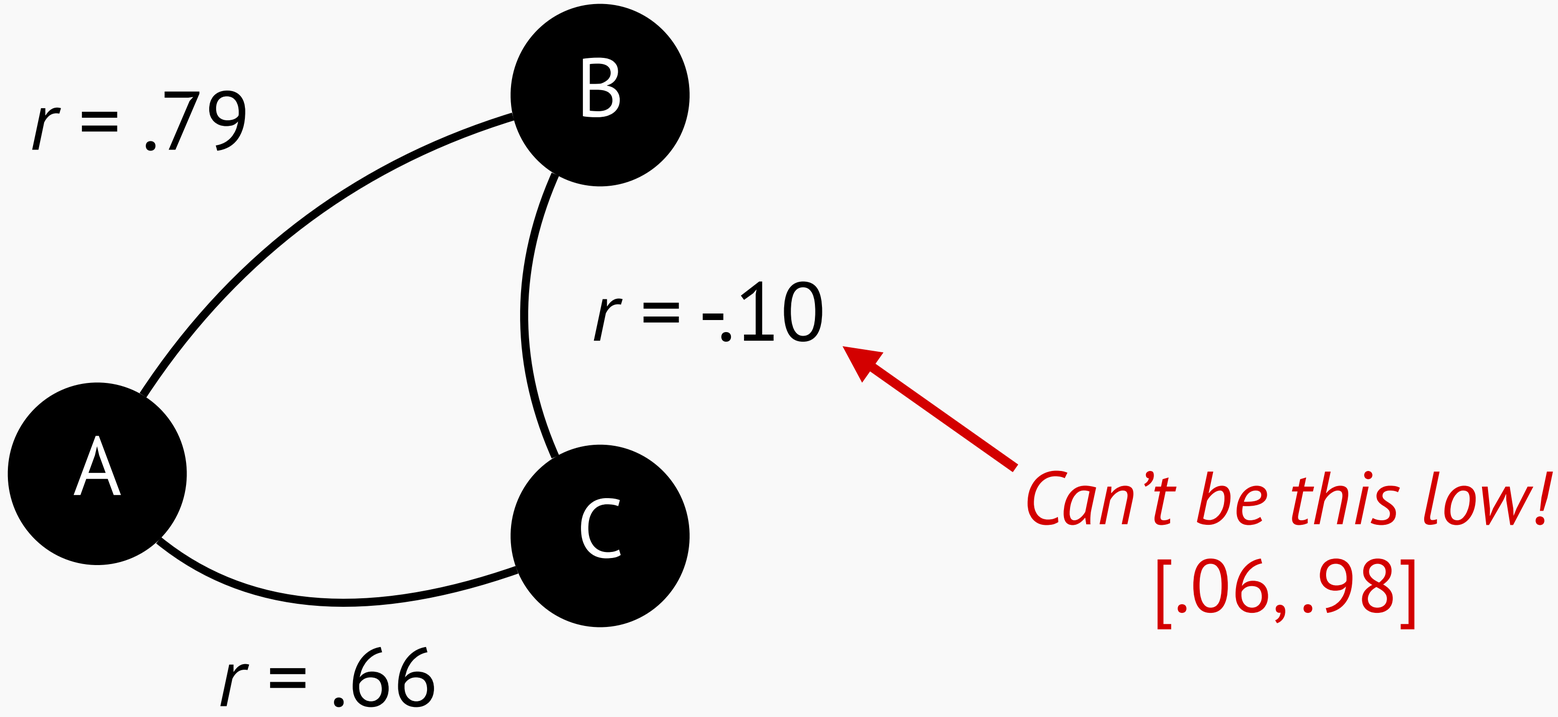
- What do you have?
  - Article, measures, data, code, the original authors' good-will?
- How can you get more?
  - Extracting additional data (plots, tables, summary statistics, asking)
- What could you check?
- How to prioritise?
  - What are the authors' claims? (Scheel, 2022)
- How to ensure reproducibility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

01.

What's wrong with these correlations?



What's wrong with these correlations?



Restrictions on correlations (Stanley & Wang, 1969)

01.

Which correlation table contains an error?

Hard to spot by eye; easier with code

	V1	V2	V3	V4	V5	V6
V1		0.32	0.34	-0.09	0.21	0.30
V2			0.45	-0.54	0.54	0.64
V3				-0.26	0.32	0.34
V4					-0.52	-0.49
V5						0.50
V6						

	V1	V2	V3	V4	V5	V6
V1		0.32	0.34	-0.09	0.21	-0.90
V2			0.45	-0.54	0.54	0.64
V3				-0.26	0.32	0.34
V4					-0.52	-0.49
V5						0.50
V6						

01.

Reusing Correlations for Validation and Regression (RECOVAR)

	V1	V2	V3	V4	V5	V6
V1		0.32	0.34	-0.09	0.21	0.30
V2			0.45	-0.54	0.54	0.64
V3				-0.26	0.32	0.34
V4					-0.52	-0.49
V5						0.50
V6						

Positive-definite

	V1	V2	V3	V4	V5	V6
V1		0.32	0.34	-0.09	0.21	-0.90
V2			0.45	-0.54	0.54	0.64
V3				-0.26	0.32	0.34
V4					-0.52	-0.49
V5						0.50
V6						

Non-positive-definite

01.

Reusing Correlations for Validation and Regression (RECOVAR)

Method development

- Statistical power?
- Corrections for rounding?
- Triangulation of problematic correlations?

Application

- Estimate prevalence of errors in correlation tables using the MetaBUS dataset (>16k publications)
  - Bosco et al. (2015)

01.

Reusing Correlations for Validation and Regression (RECOVAR)

	V1	V2	V3	V4	V5	V6
V1		0.32	0.34	-0.09	0.21	0.30
V2			0.45	-0.54	0.54	0.64
V3				-0.26	0.32	0.34
V4					-0.52	-0.49
V5						0.50
V6						

Fit regressions using correlation tables rather than data

- Reproducibility
- Robustness
- Probe researchers' implicit causal assumptions

$$V1 \sim V2 + V3 + V4$$

UNIVERSITÄT BERN

A master’s degree course in error detection

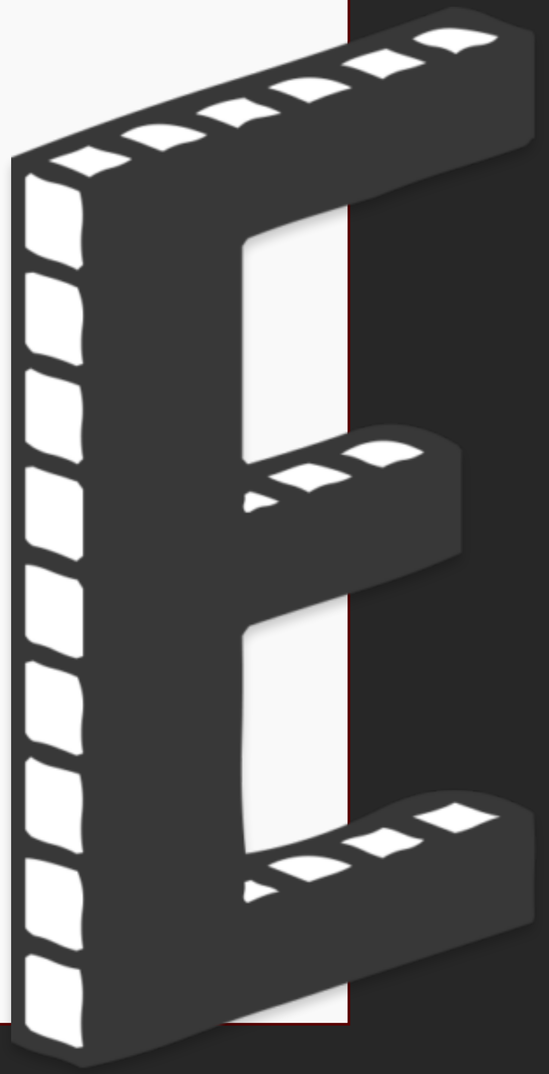
xxxxxxxxxxxxxxxxxxxxxxxxxx

02.

02.

Conspicuous absence of error detection in the curriculum

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



02.

‘Normalise’ is a verb

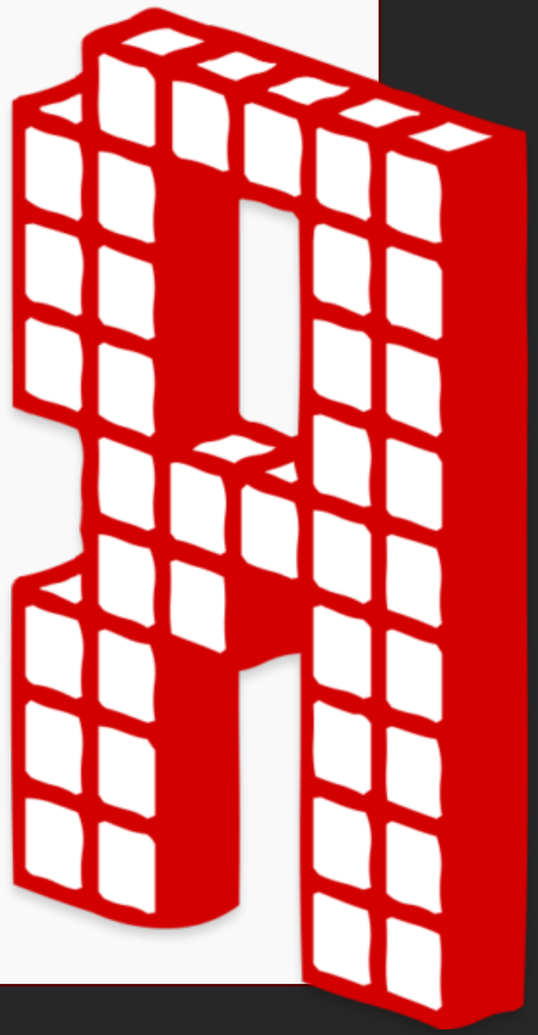
Masters course:

“Estimating the credibility of past research”

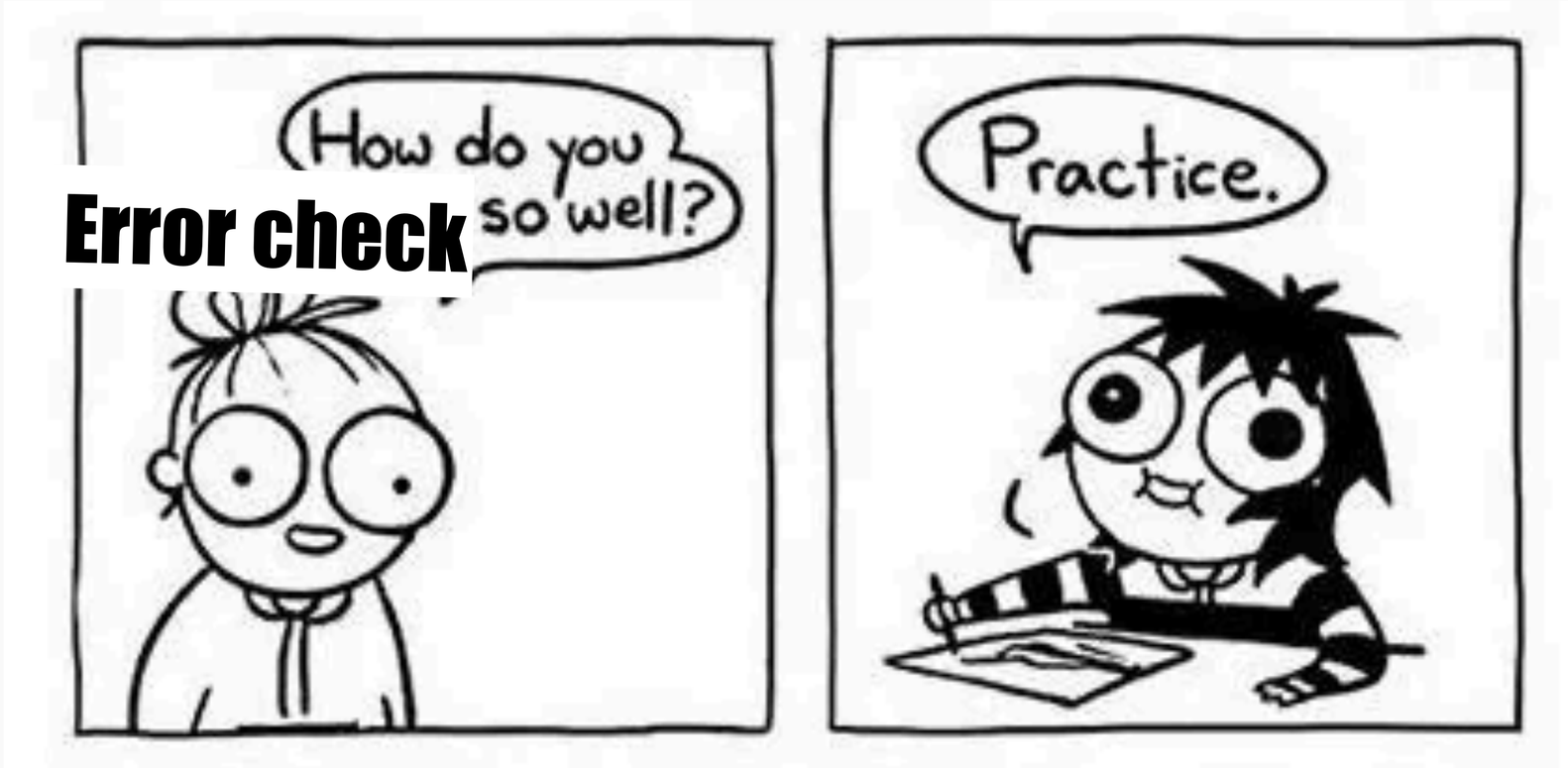
(@UniBe from Fall ’23)

Science as an intensely human & fallible activity

Organised skepticism



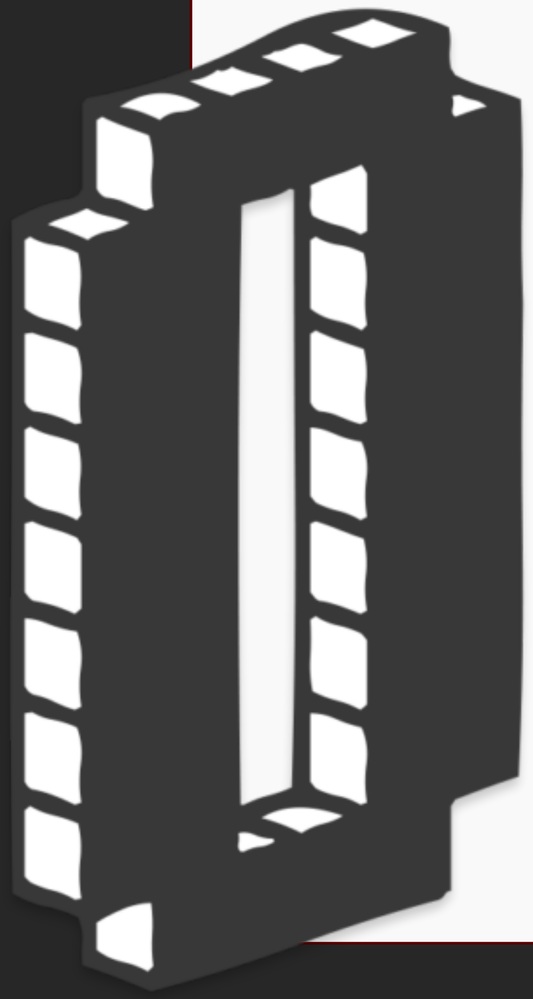
02.



Error check

Syllabus

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



Not merely information, practice!

02.

Problems encountered

Permission to critique

xxxxxxxxxx

Language of critique

Exposure & Practice!

UNIVERSITÄT BERN

ERЯOR

A bug bounty program for science

xxxxxxxxxxxxxxxxxxxxxxxxxx

03.

250,000 CHF

(€260k, $285k)

fund

Check

100

published articles
for errors

Pay authors & reviewers

+ bonus reward

contingent on errors found

	Base	**Bonus reward**
Author	CHF 250	Up to 250 CHF
Reviewer	c.250 – 1,000 CHF Depending on expected effort	Up to 2,500 CHF

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

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**

Candidate articles

- Selected for ‘importance’:
  Would matter if it contained errors

- Recent (c.10 years)

Obtain author consent

- Suggestions welcome
- Especially self-nominations

Match with reviewer

- Not randomly sampled

- Attention paid to gender balance etc.

- Consent to have their publicly available scientific output scrutinised by peers + get paid for it

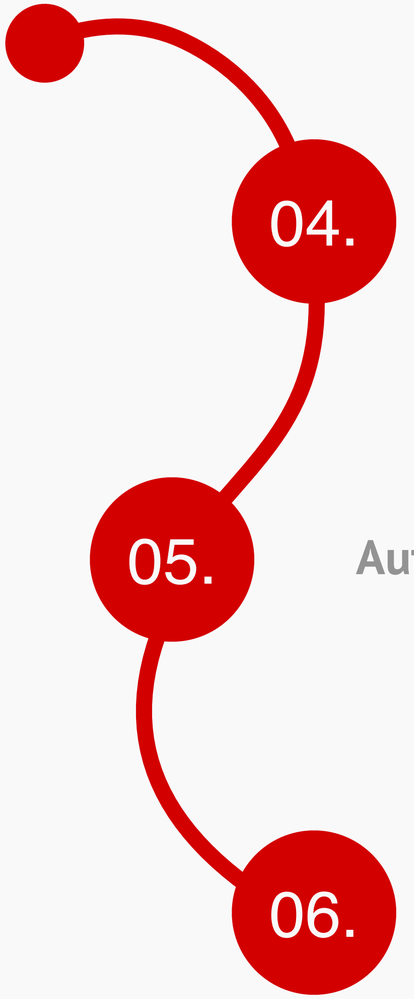
- Pre-commit to:
  - Sharing materials
  - Answering reasonable questions
  - Professional conduct
  - Acting on any recommendations

Candidate articles

Obtain author consent

Match with reviewer

- Volunteers welcome!
- Non-anonymous
- Can ask author reasonable questions to aid their work
- Pre-commit to professional conduct



Error report

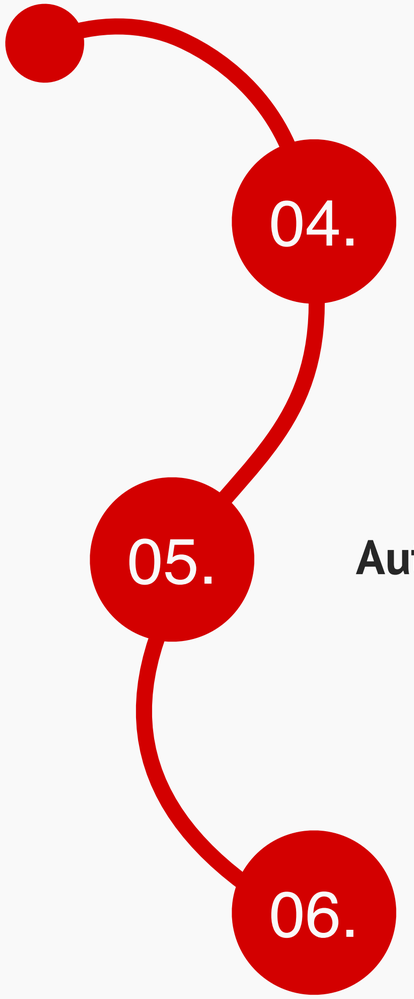
- Checked by recommender
- Claims must be substantiated

Author response

- Recommender can ask to modify/ correct/edit
  - Move material from ‘Errors’ section to ‘Broader critiques’ section

Decision, recommendation, & reward payments

- Published on error.reviews website
  - CC-By 4.0 licence

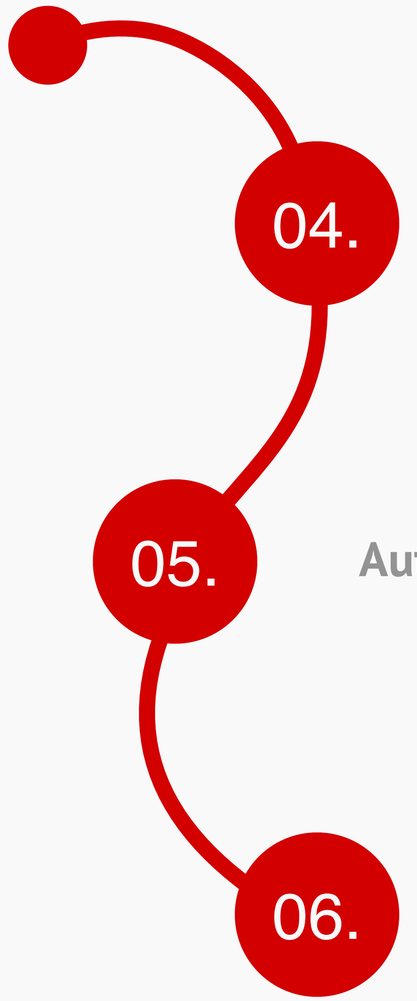


Error report

Author response

Decision, recommendation, & reward payments

- Optional, timely
- Recommender can ask to modify/ correct/edit
- Published on error.reviews website
  - CC-By 4.0 licence



Error report

Author response

Decision, recommendation, & reward payments

- Recommender makes a decision & any associated recommendation
- Possibly in conjunction with the advisory board

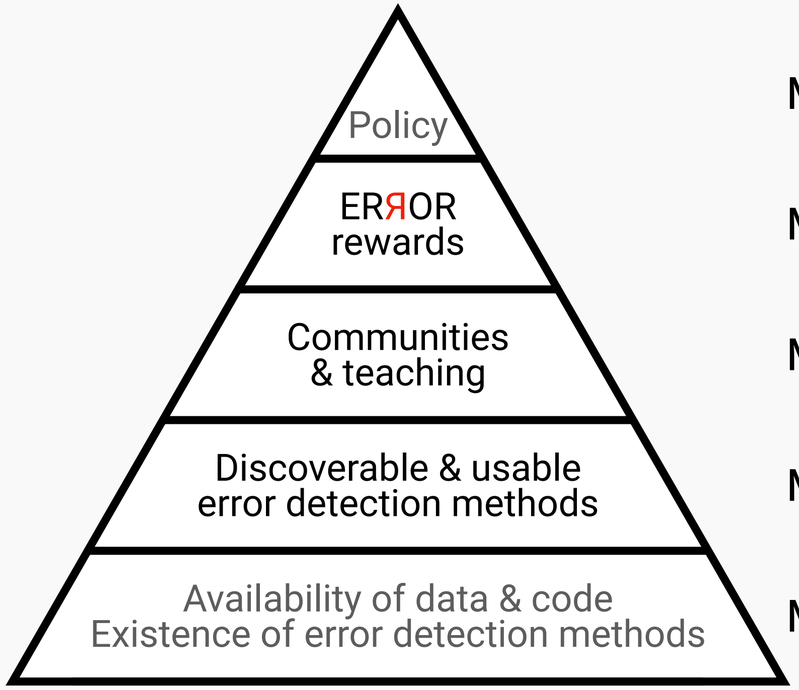
Recommendations

Decision	Recommendation
No errors	No additional action beyond publication of error report on the error.reviews website.
Minor errors	Authors to appropriately recognise these errors in future discussions of the article. *Probably most research!*
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.
Moderate errors	Correction notice (minor)
Major errors	Correction notice (major) / may warrant an expression of concern
Severe errors	Retraction

Bonus rewards

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 errors	Authors to appropriately recognise these errors in future discussions of the article. *Probably most research!*	250 CHF 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.	250 CHF to the reviewer
Moderate errors	Correction notice (minor)	500 CHF to the reviewer
Major errors	Correction notice (major) / may warrant an expression of concern	1000 CHF to the reviewer
Severe errors	Retraction	2500 CHF to the reviewer

Our strategy to increase error checking, reporting, & correction



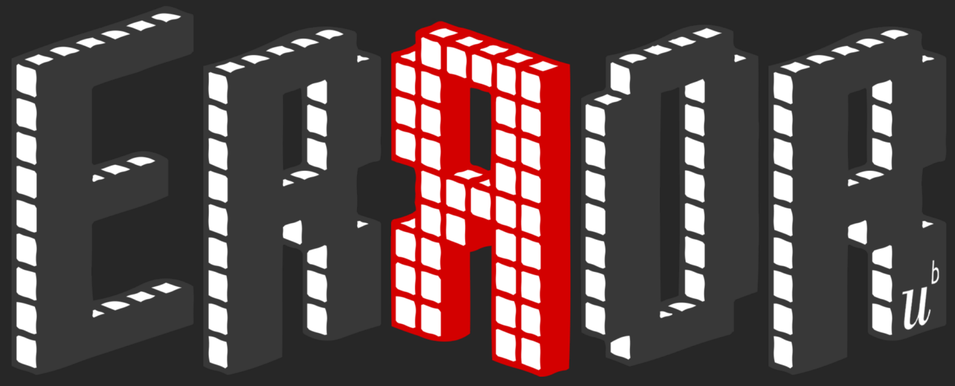
Make it required

Make it rewarding

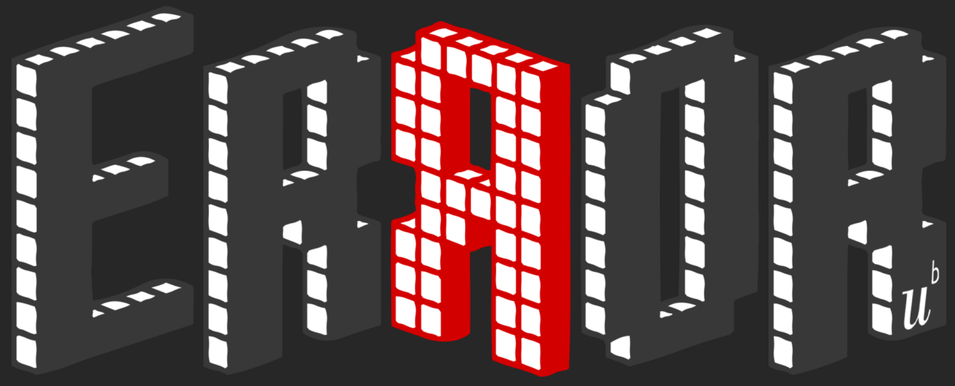
Make it normative

Make it easy

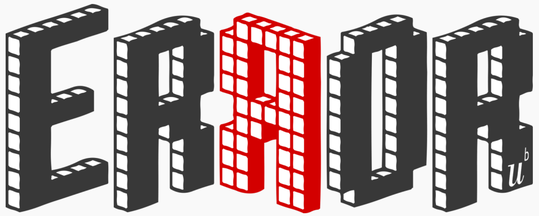
Make it possible



0 / 3	10	–	0 / 250'000 CHF
Reviews Completed / Pending	Invited Papers	Papers with Errors	Total Payouts



error.reviews



Contact

ian.hussey@unibe.ch
Twitter: @ianhussey

https://error.reviews
error-reviews.psy@unibe.ch
Twitter: @error_reviews

UNIVERSITÄT BERN

xxxxxxxxxxxxxxxxx

Acknowledgements

Collaborators

- Malte Elson
- Ruben Arslan
- Frank Bosco
- Beth Clarke
- Alex Holcombe
- Imran Kadolkar

ERROR authors & reviewers to date

- Dale Barr
- Lisa DeBruine
- Emma Henderson
- Gordon Pennycook and colleagues
- Daniel Simons
- Jan Wessel

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