2023-05-02

Responsible Research Assessment

A practical recommendation for the evaluation of research quality beyond *h*-index and journal impact factors

Felix Schönbrodt, Anne Gärtner, Daniel Leising

Prof. Dr. Felix Schönbrodt
Ludwig-Maximilians-Universität
München

OSC
LMU Open Science Center

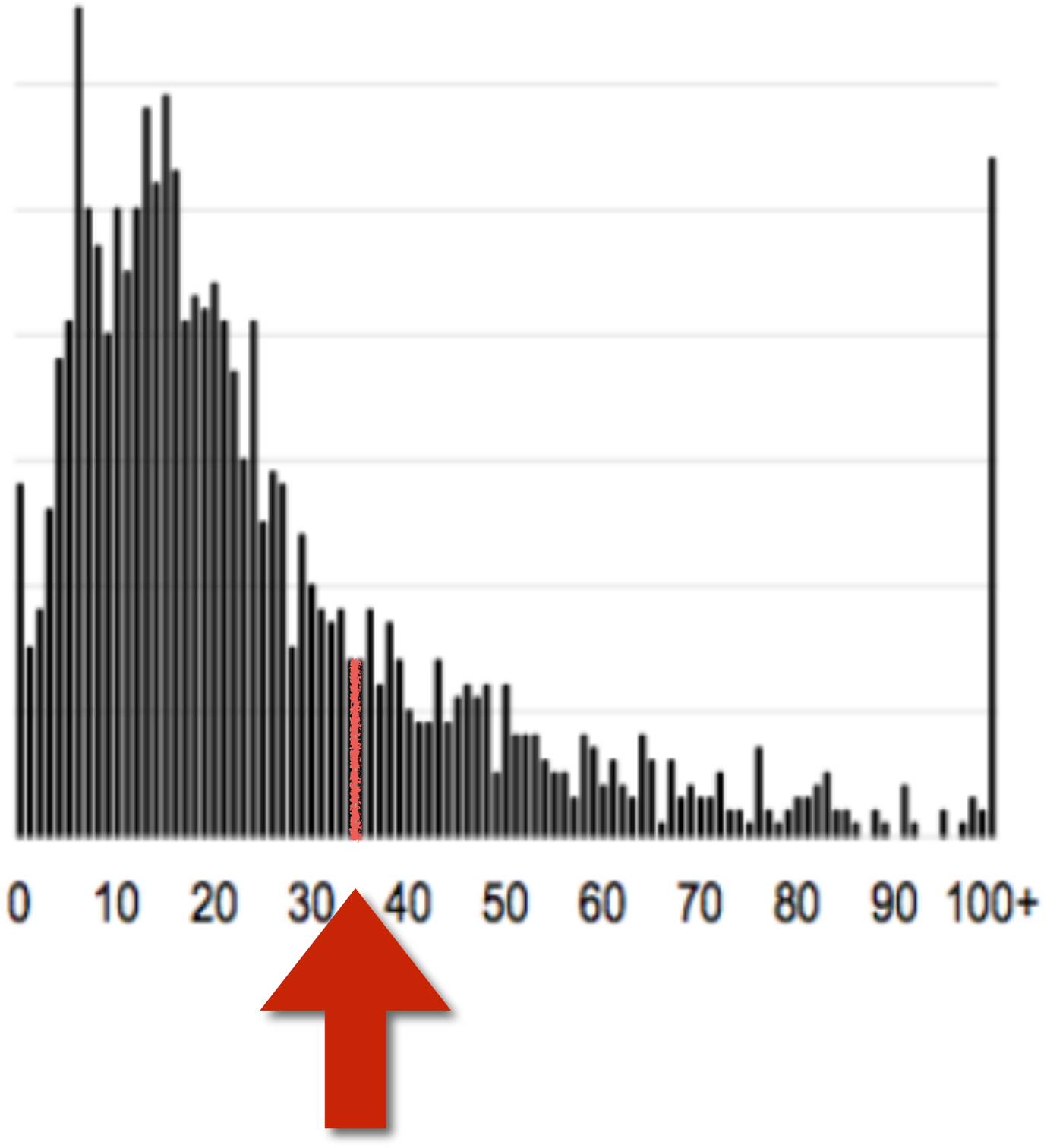
www.osc.uni-muenchen.de
www.researchtransparency.org
@nicebread@scicomm.xyz

Part I:

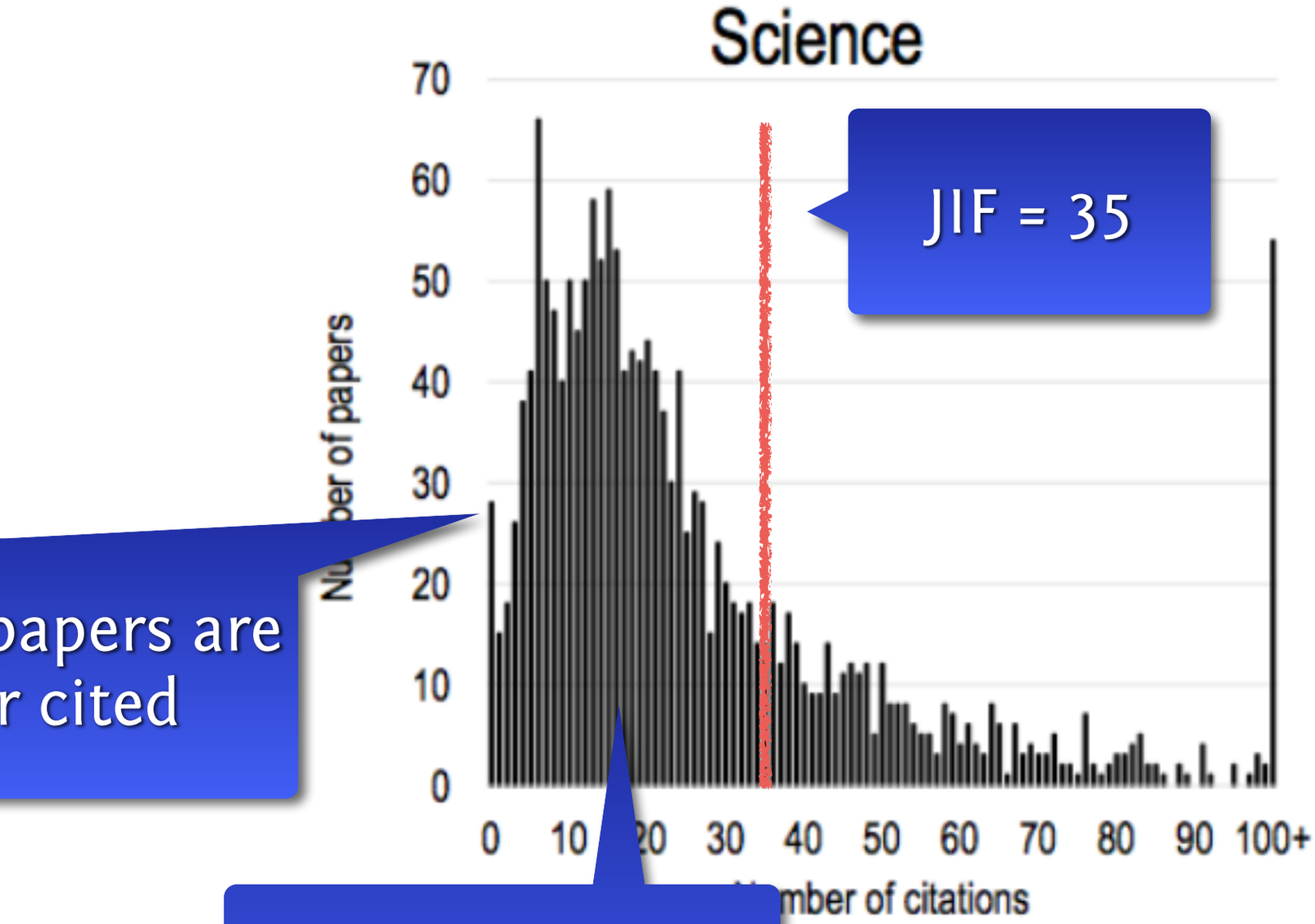
Problems with the current approach

Thesis 1:

Our current indicators for scientific quality do a bad job.


0
10
20
30
40
50
60
70
80
90
100+

Journal Impact Factor (JIF)



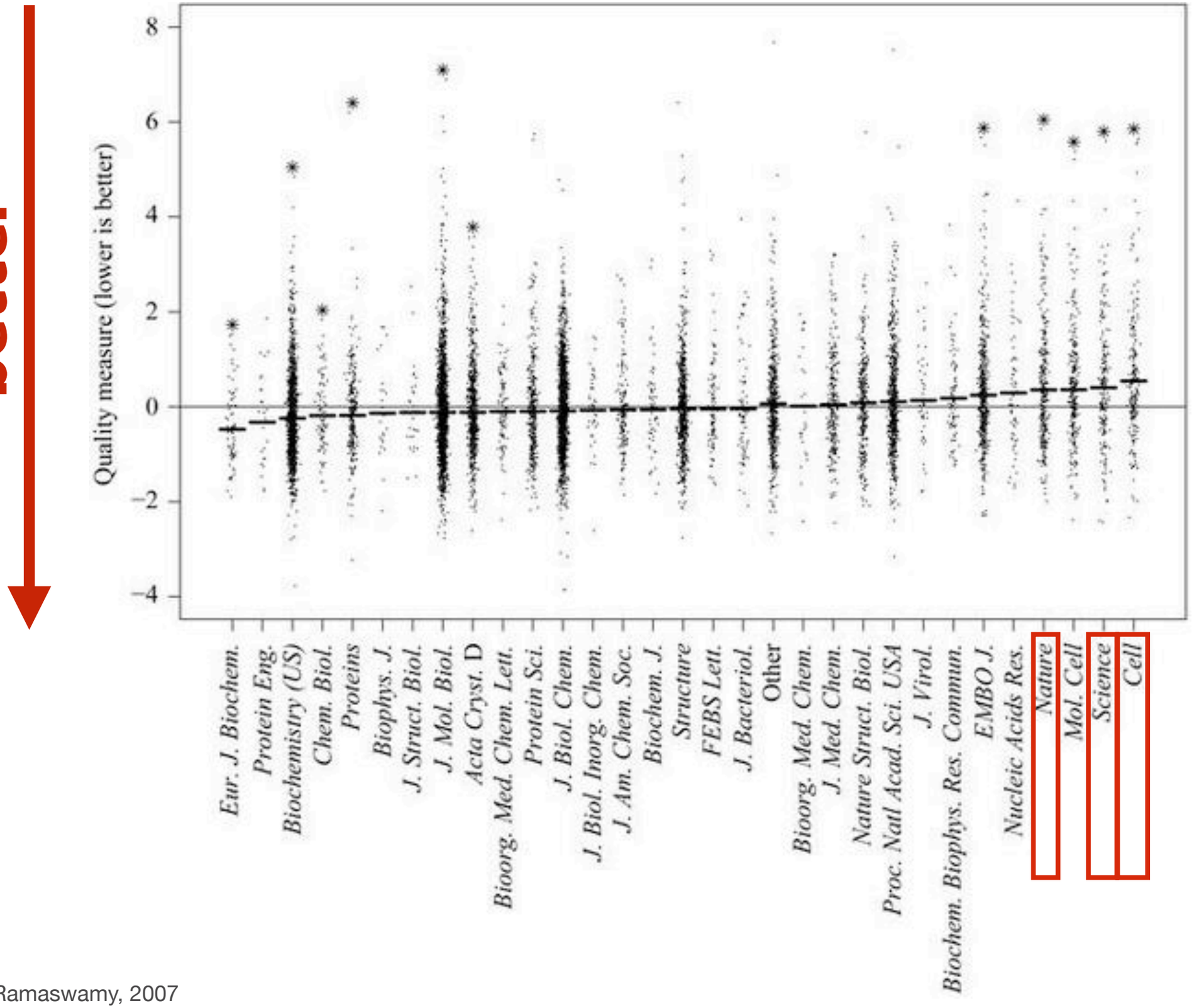
Lariviere, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., u. a. (2016). A simple proposal for the publication of journal citation distributions. bioRxiv, 062109. doi:10.1101/062109

Journal Impact Factor (JIF)

- Objective quantification of cristallograhic quality: higher JIF ➔ less quality (Brown and Ramaswamy, 2007)

Für Überblick, siehe Brembs, Button, & Munafò (2013)

Journal Impact Factor (JIF)



Brown and Ramaswamy, 2007

Journal Impact Factor (JIF)

- Objective quantification of cristallograhic quality: higher JIF → less quality (Brown and Ramaswamy, 2007)
- *Positive* relationship between JIF and objective errors of gene names in Excel sheets (Ziemann, Eren, & El-Osta, 2016)
- *Positive* relationship between JIF and the frequency of retractions (Brembs, Button, & Munafò, 2013)
- *Negative* relationship between JIF and probability of replication (Dougherty & Horne, 2022)
- In vivo animal experimentation studies are less randomized and report more conflicts of interest in higher ranking journals (Macleod et al., 2015)

For an overview, see Brembs, Button, & Munafò (2013); Brembs, 2018; Dougherty & Horne, 2022

„Double blind peer review is the hallmark of scientific quality control“

Reliability for single case diagnostics

How well do reviewers agree in their assessment of a paper?

➙ interrater agreement

angeben. Reliabilitäten um .70 oder weniger gelten als unbefriedigend und sind für die Einzelfalldiagnostik nicht geeignet. Ab Werten von .80 kann man von einer für die Einzelfalldiagnostik akzeptablen Reliabilität sprechen. Tests zur

Goal:

ICC > .80

Table 6.3: Altman's Kappa Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very Good

kappa > .60

Krohne, H. W., & Hock, M. (2015). Psychologische Diagnostik: Grundlagen und Anwendungsfelder. Kohlhammer Verlag.

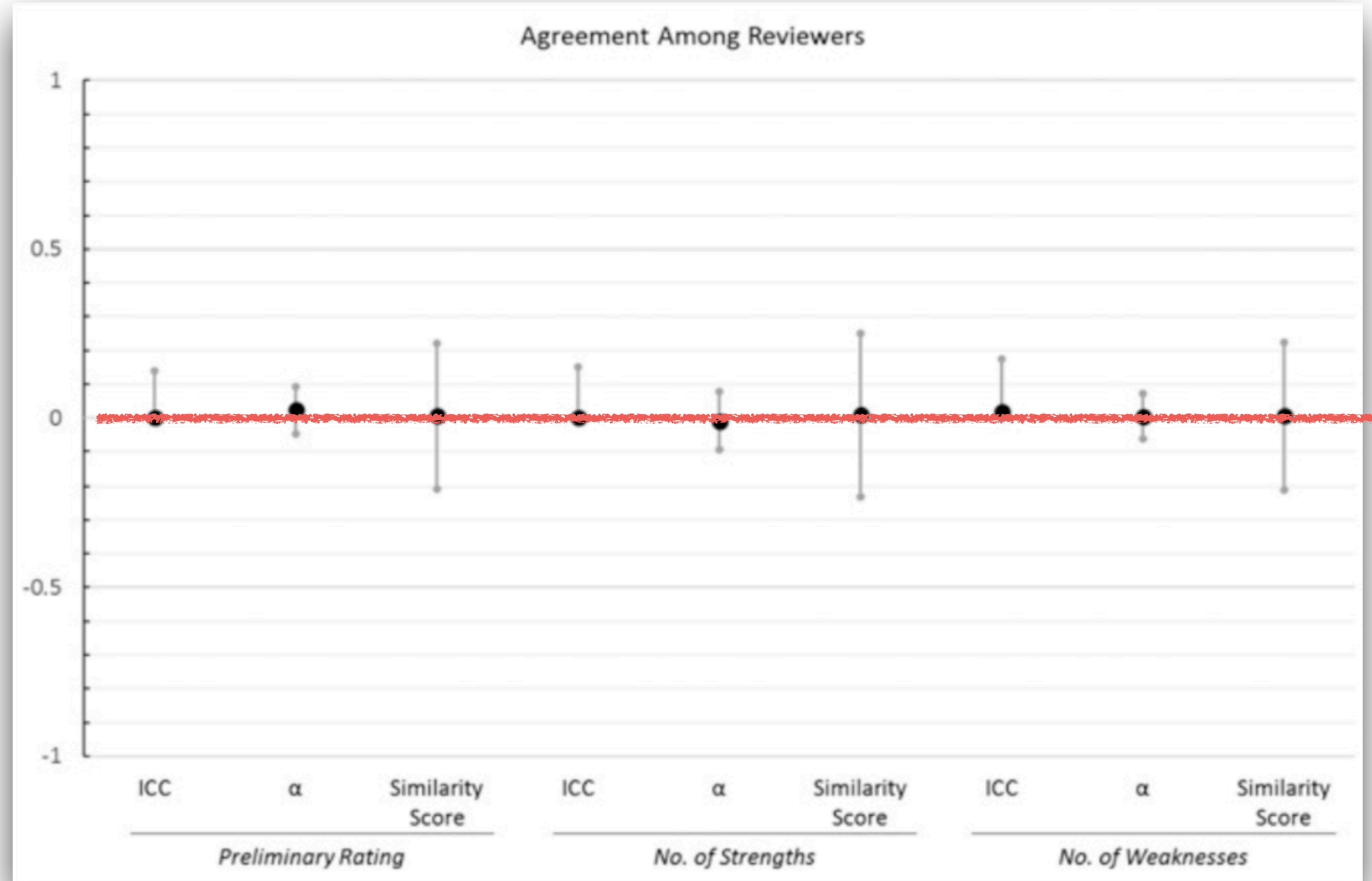
Peer review

- Meta-analysis of reviewer agreement (*k*=48, 19,443 manuscripts): ø ICC = .34, kappa = .17 Bornmann, Mutz, Daniel (2010)

for an overview, see Osterloh, M., & Kieser, A. (2015).

Peer review

0 (zero) agreement



Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., u. a. (2018). Low agreement among reviewers evaluating the same NIH grant applications. Proceedings of the National Academy of Sciences, 201714379. doi:10.1073/pnas.1714379115

Peer review

- Meta-analysis of reviewer agreement (*k*=48, 19,443 manuscripts): ⌀ ICC = .34, kappa = .17 Bornmann, Mutz, Daniel (2010)

- „Agreement about shared values“ ≠ „agreement about true value“
  ➞ Correlation with „true value“ <= 1
  ➞ Estimate correlation of reviewer's assessment with „true value“ of a paper: *r* = .09 – .27 (mean *r* = .18; explained variance: 3%) Starbuck (2004)

- Decisions highly dependent on a (random) selection of reviewers ➞ Lottery Bornmann & Daniel (2009)

for an overview, see Osterloh, M., & Kieser, A. (2015).

„When I divide the week's contribution into two piles – one that we are going to publish and the other we are going to return – I wonder whether it would make any real difference to the journal or its readers if I exchanged one pile for another".

Sir Theodore Fox in *Lancet*, 1965

SHIN, J. C., Toutkoushian, R. K., & Teichler, U. (2011). University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education. Springer Science & Business Media, p. 151.; Zitat erstmalig gesehen bei Präsentation von Margit Osterloh

Peer review

- Meta-analysis of reviewer agreement (*k*=48, 19,443 manuscripts): ⌀ ICC = .34, kappa = .17 Bornmann, Mutz, Daniel (2010)

- „Agreement about shared values“ ≠ „agreement about true value“
  ➞ Correlation with „true value“ <= 1
  ➞ Estimate correlation of reviewer's assessment with „true value“ of a paper: *r* = .09 – .27 (mean *r* = .18; explained variance: 3%) Starbuck (2004)

- Decisions highly dependent on a (random) selection of reviewers ➞ Lottery Bornmann & Daniel (2009)

- Summary: Pre-publication peer review in the current system has some value as *feedback* (to improve a manuscript), but limited value as *quality control*; very inefficient.

- Silver lining: Agreement is higher at the low quality end (Cicchetti, 1991)

for an overview, see Osterloh, M., & Kieser, A. (2015). See also

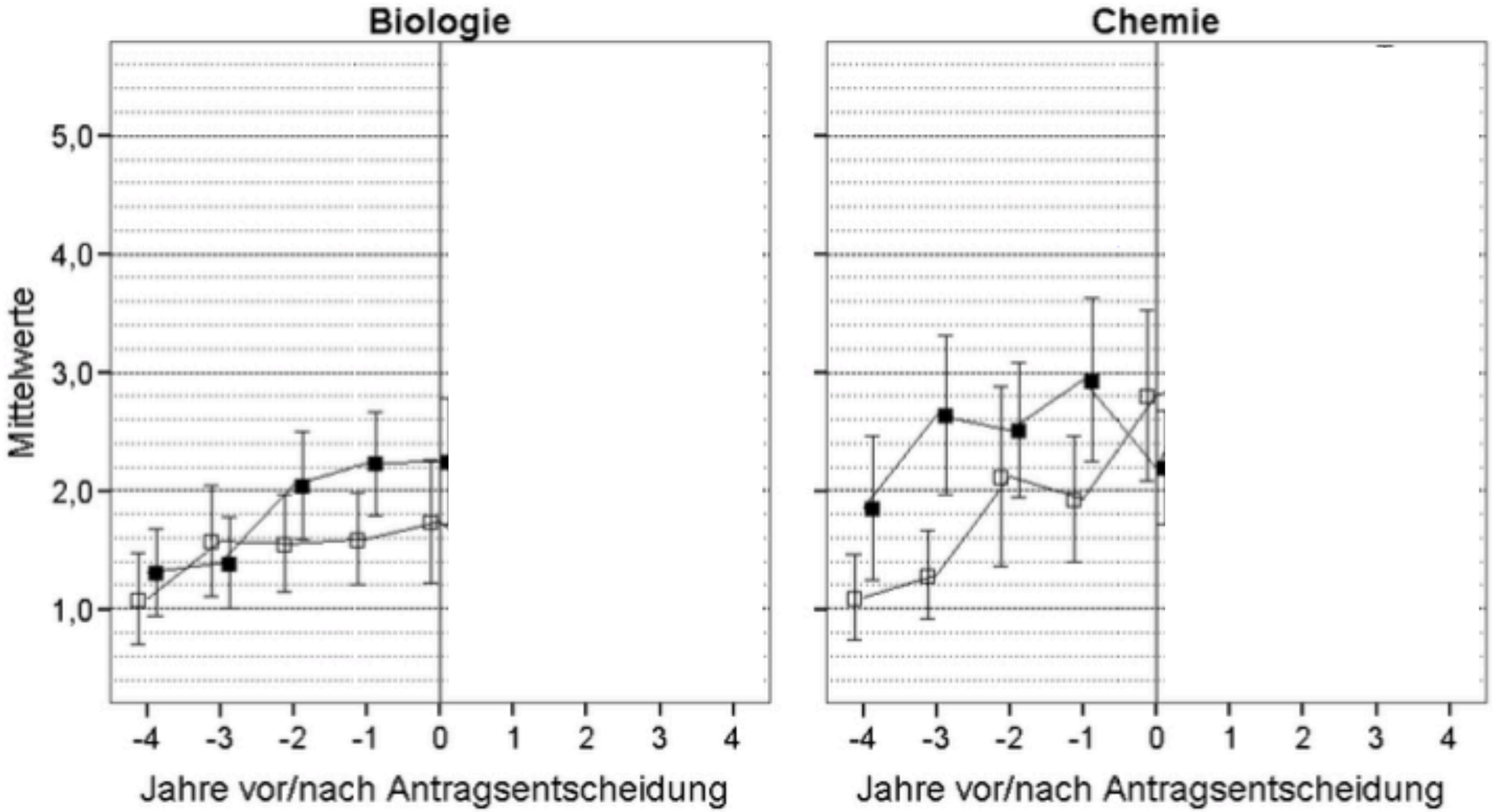
Does peer review select better researchers?

A case study about the prestigious Emmy-Noether-Programme in Germany

„The Emmy Noether Programme gives **exceptionally qualified** early career researchers the chance to qualify for the post of professor at a university by leading an independent junior research group for a period of six years.“

Emmy-Noether-Programm

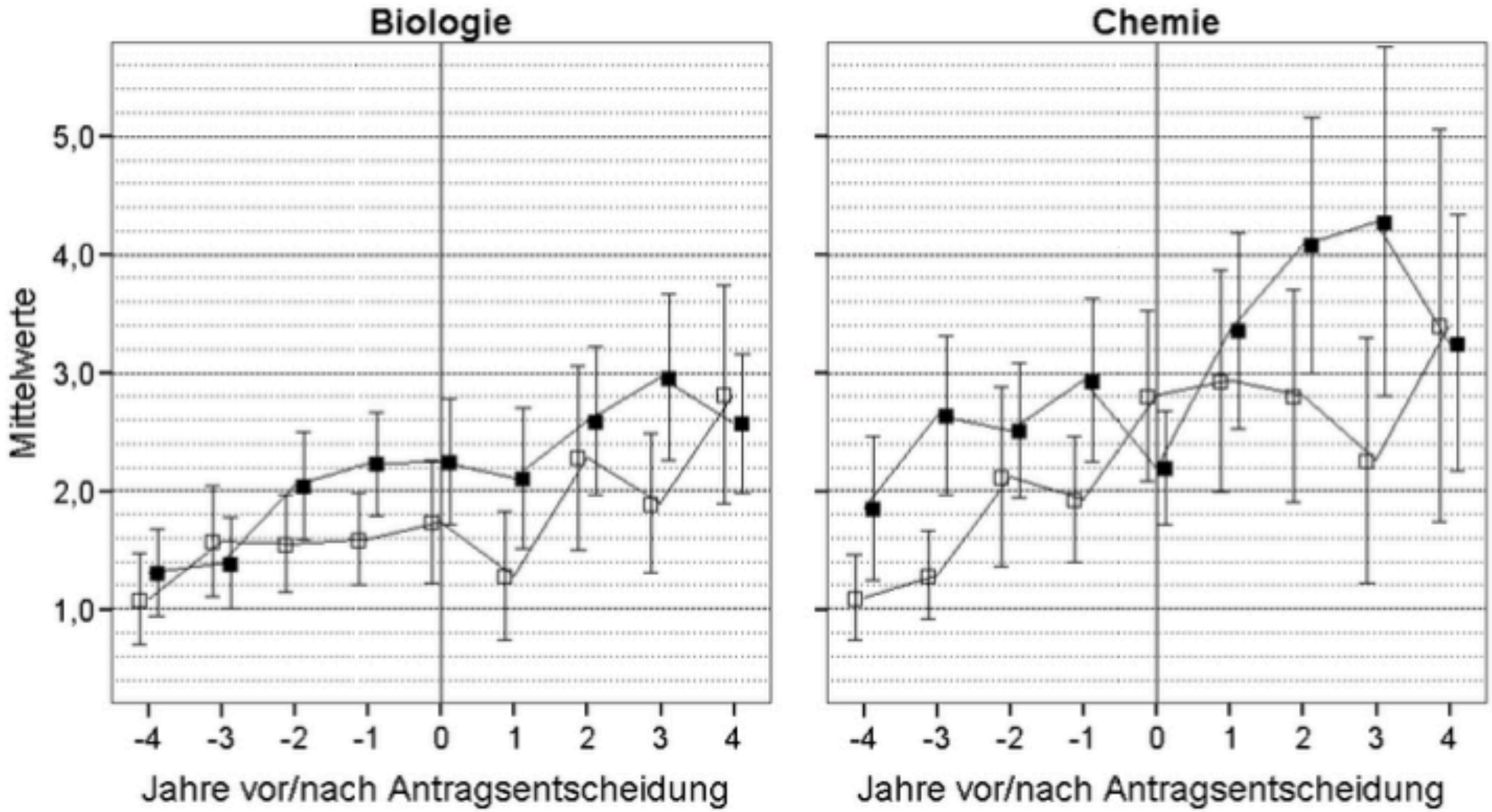
Number of publications



Böhmer et al (2008); Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & Ins, von, M. (2008); Neufeld, J. (2016).

Emmy-Noether-Programm

Number of publications



Fehlerbalken = 95% ci

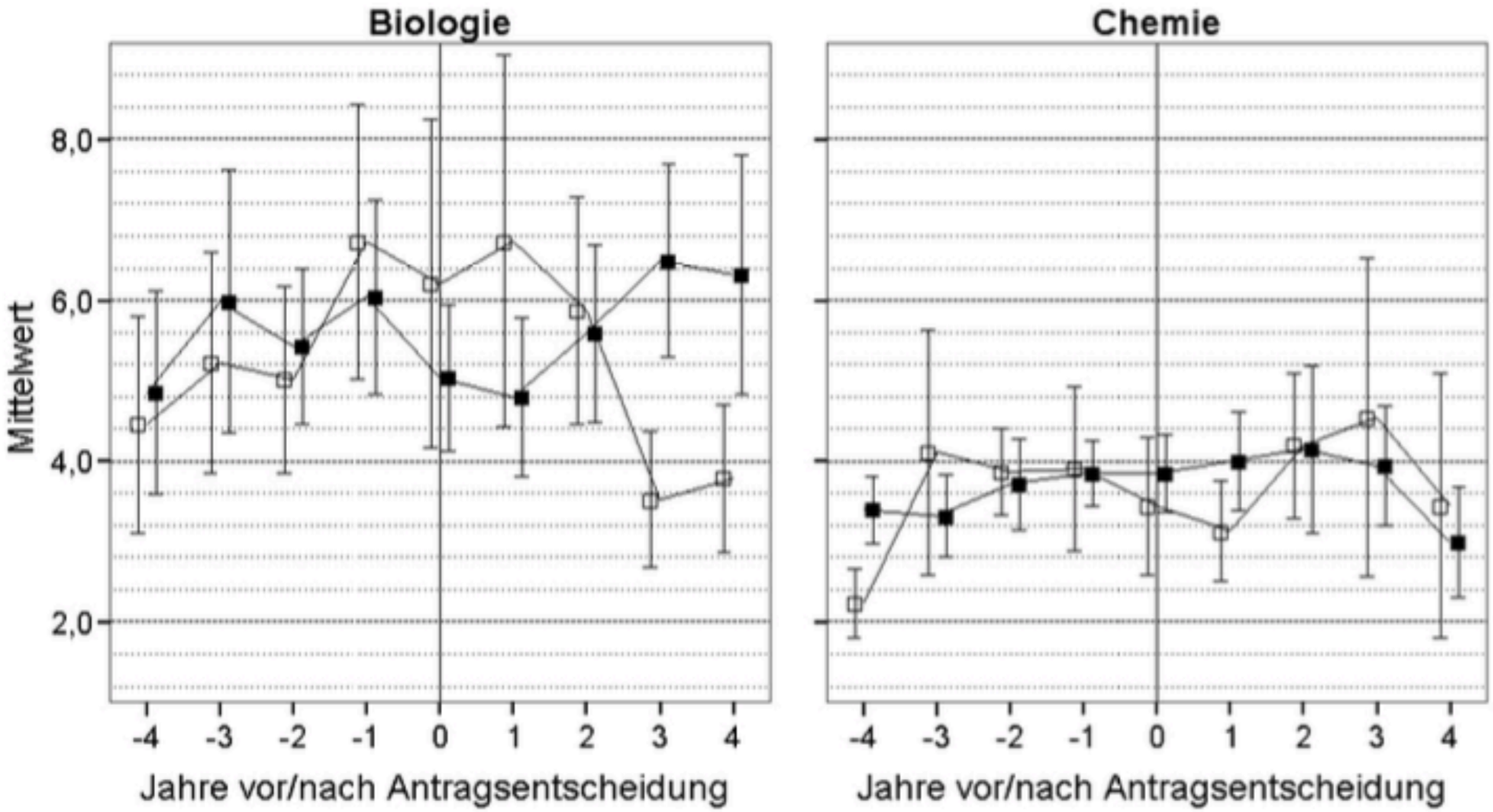
reiected

accepted → receive ~1.500.000 €

Böhmer et al (2008): Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & Ins, von, M. (2008); Neufeld, J. (2016).

Emmy-Noether-Programm

Average JIF



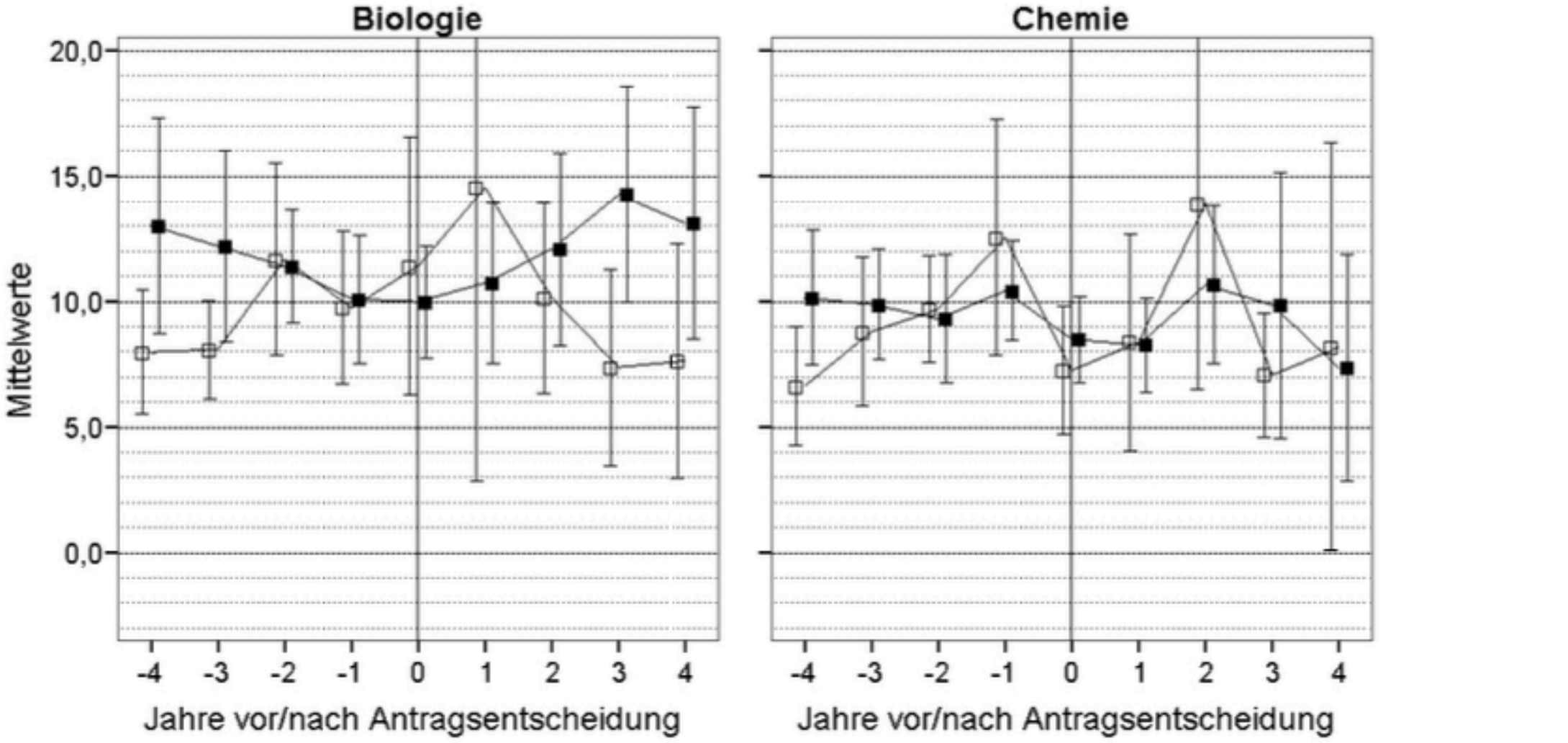
rejected

accepted → receive ~1.500.000 €

Böhmer et al (2008): Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & Ins, von, M. (2008); Neufeld, J. (2016).

Emmy-Noether-Programm

Citations per Paper



Böhmer et al (2008); Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & Ins, von, M. (2008); Neufeld, J. (2016).

Emmy-Noether-Programm

"Taken together, the bibliometric results show remarkably small differences between funded and rejected applicants (prior to funding). Moreover, these small differences are not increased by the fact that one of both groups gets the funding of the Emmy Noether program* and the other doesn't."

**1 - 1.5 million €*

Böhmer et al (2008)

Thesis 2:

Our current incentives foster bad science.

Richard Horton,
Editor von *The Lancet*

Much of the scientific literature, perhaps half, may simply be untrue.

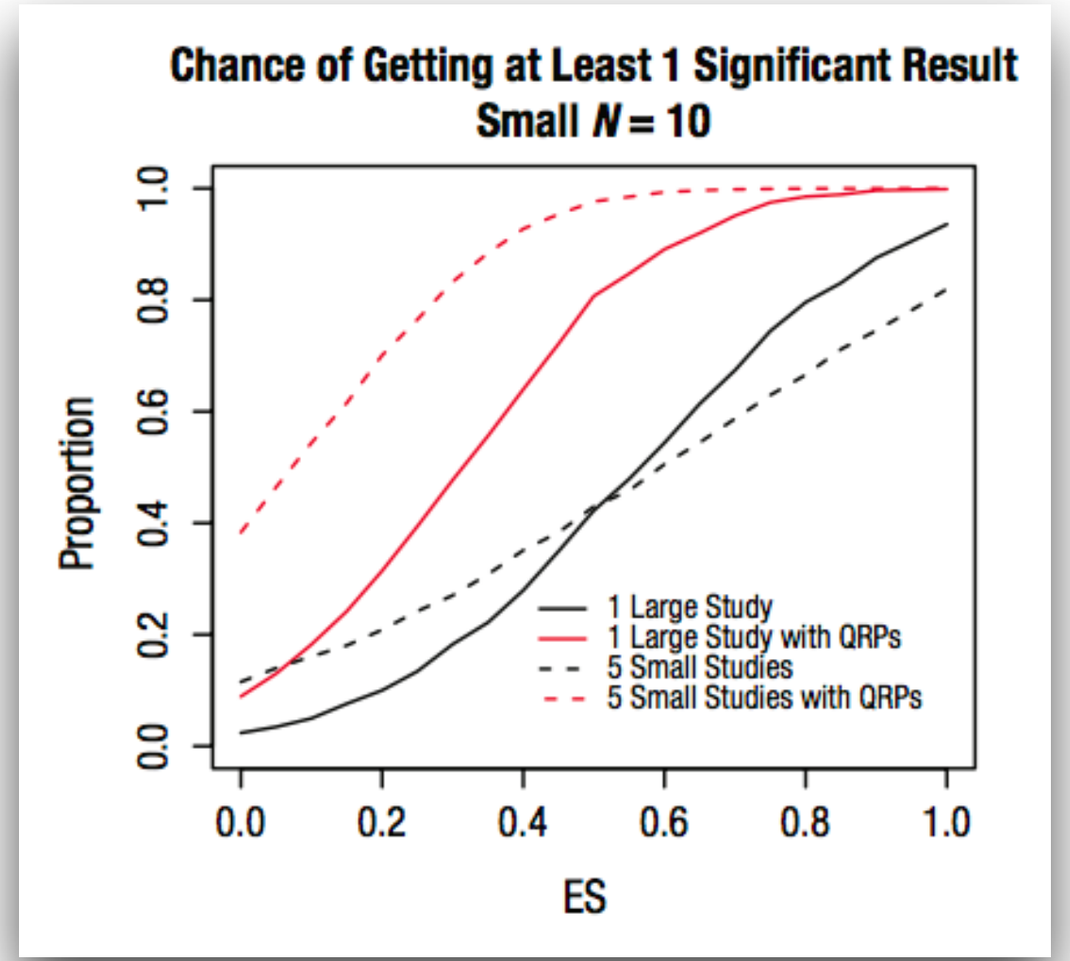
Part of the problem is that no one is incentivised to be **right**.

Quantity, not quality

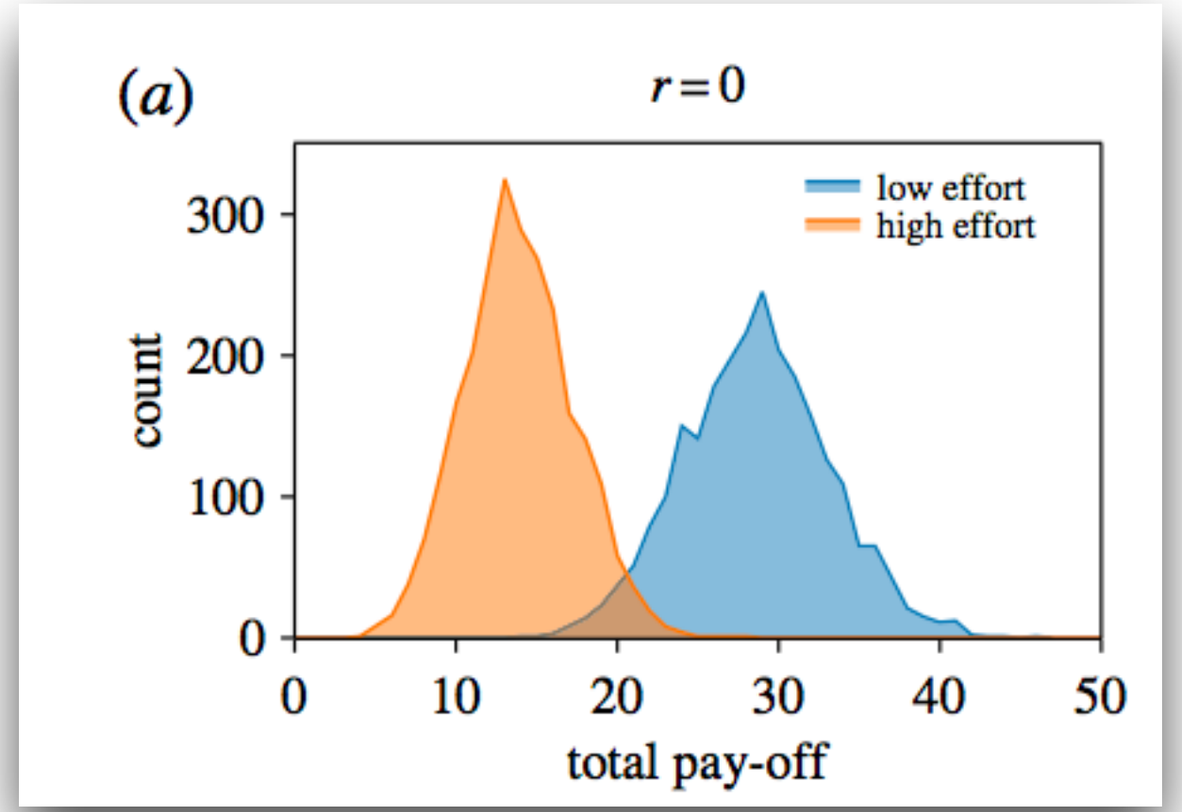
Actual (not desired) relevance at professorship hiring committees:	Rank
Number of peer-reviewed publications	1
Fit of research profile to the advertising institution	2
Quality of research talk	3
Number of publications	4
Volume of acquired third-party funding	5
Number of first authorships	6
…	…

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? Psychologische Rundschau, 67(4), 250–261. http://doi.org/10.1026/0033-3042/a000335

„The rules of the game“



„Evolution of bad science“



Ideal strategy for a high quantity of publications:
small *n* + many studies + questionable research practices (QRPs), such as *p*-hacking

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. Perspectives on Psychological Science, 7(6), 543–554.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384–17.

„It's impossible to win the Tour de France without doping."

Lance Armstrong

https://www.lemonde.fr/sport/article/2013/06/28/avant-le-tour-lance-armstrong-brise-le-silence_3438032_3242.html

Quotes from early career researchers

… collected anonymously for a Q&A session in a workshop on open science.

With which mindset do we conduct research - is our goal to find new truths or to be successful by „confirming" our hypotheses?

What would be a good **balance** between Open Science and pursuing a career in science?
Can you only really afford doing Open Science with tenure?

Open Science makes me transparent, but also very vulnerable. Is it really worth it?

My contract is limited to two years – although it would be nice to publish the data, I have no time to do it. I rather have to churn out another publication.

➞ felt contradiction between „good research"/„open research" and „having a career in science"

Doing „good research“ vs. „having a career in science“?

© KC Green

It is the task of the senior researchers
- those sitting in committees and making the rules of hiring, tenure and promotion -
to solve that dilemma for the early career researchers!

Part II:
How to realign good scientific practice and incentive structures

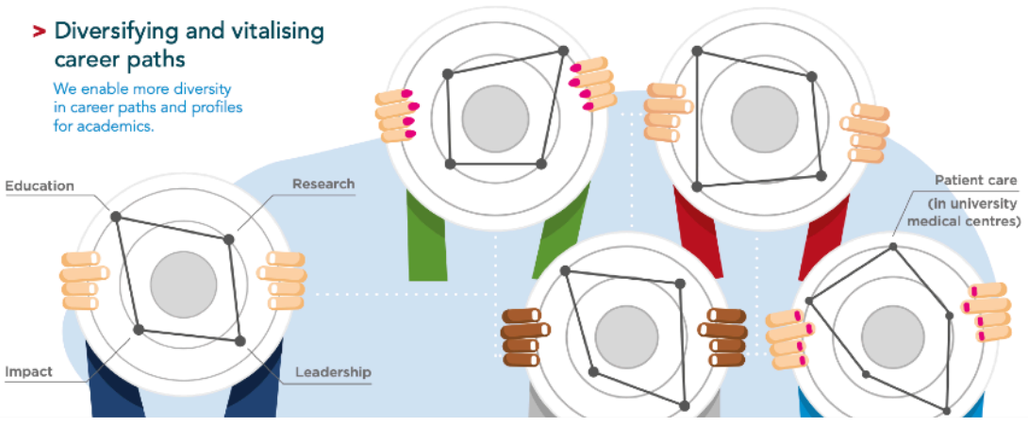
DORA

> The aim is for research to be evaluated based on its intrinsic merits rather than on the number of publications and where these are published.
>
> *Scoping report on research assessment. European Commission*

Paris Call on Research Assessment

Room for everyone's talent

towards a new balance in the recognition and rewards of academics



LERU

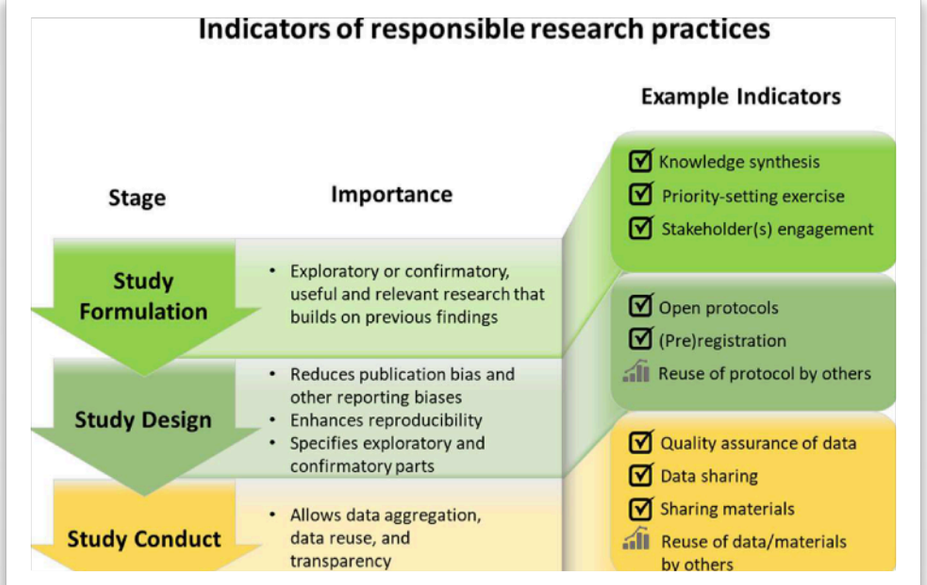
PUSHING THE FRONTIERS OF INNOVATIVE RESEARCH

A Pathway towards Multidimensional Academic Careers

A LERU Framework for the Assessment of Researchers

LERU position paper
January 2022

Hong Kong Principles



European Commission

Towards a reform of the research assessment system

Scoping Report

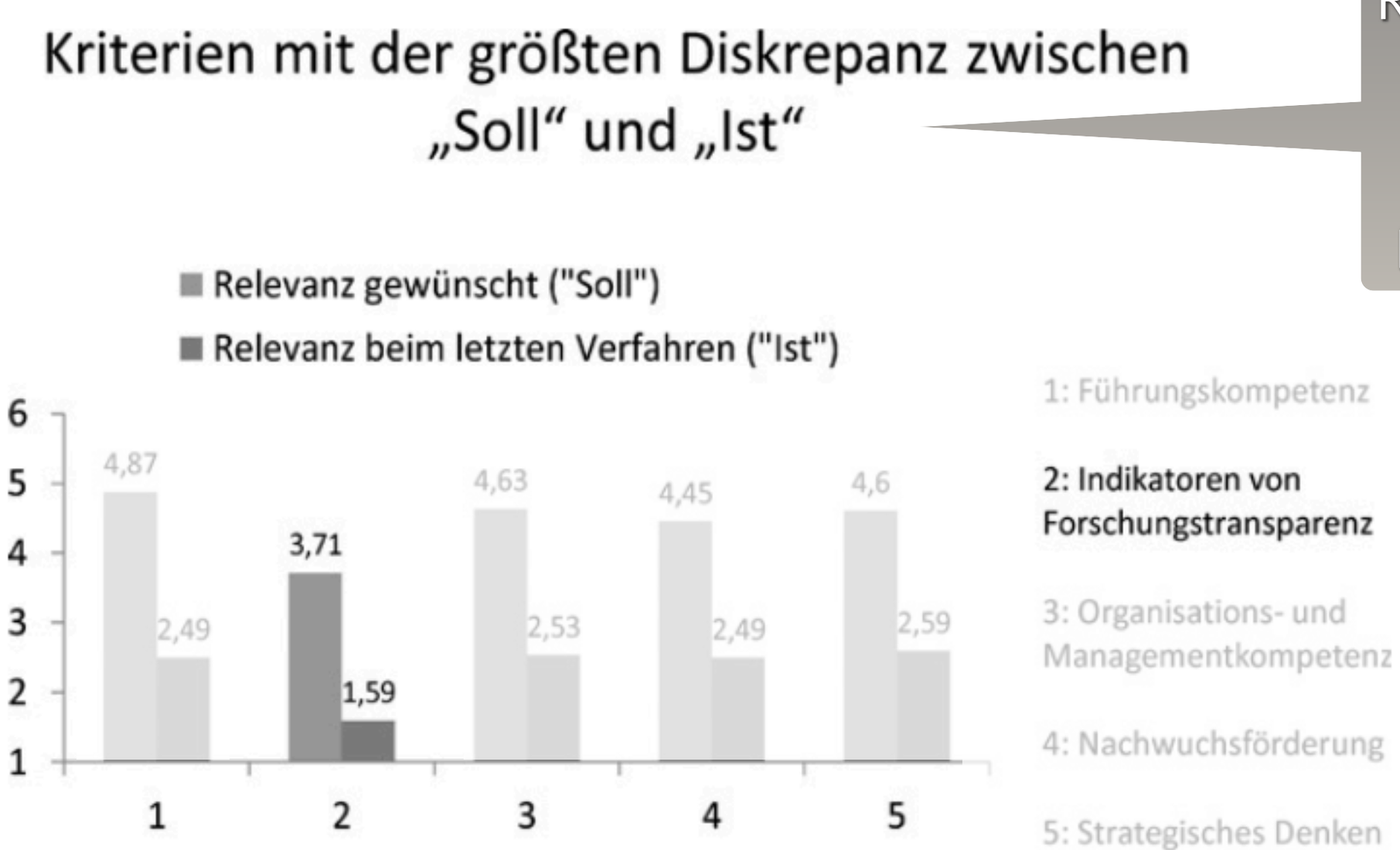
Quantity, not quality

Actual (not desired) relevance at professorship hiring committees:	Rank
Number of peer-reviewed publications	1
Fit of research profile to the advertising institution	2
Quality of research talk	3
Number of publications	4
Volume of acquired third-party funding	5
Number of first authorships	6
…	…
Quality assessment of the best three publications	17
…	…
Indicators of research transparency	**41 (of 41)**

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? Psychologische Rundschau, 67(4), 250–261. http://doi.org/10.1026/0033-3042/a000335

Quality, not quantity

Indicators with the largest discrepancy between „desired“ and „actual“: Researchers **want** to have indicators of research transparency in hiring committees!



Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? Psychologische Rundschau, 67(4), 250–261. http://doi.org/10.1026/0033-3042/a000335

Responsible Research Assessment: A proposal for professorship hiring committees

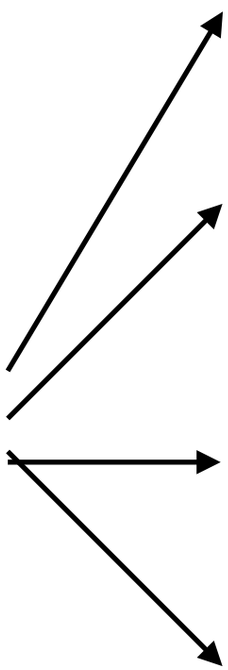
Effectiveness of any new assessment scheme

=

Quality ⟶ More **valid indicators** & assessment procedures that combat biases

x

Acceptance



First-mover / collective action problem ➙
CoARA is the best shot so far …

Transparency of requirements & **reproducibility** of indicators: No more proprietary black box algorithms

Efficiency in the hiring committee:
Can it handle 100+ applicants?

Cultural change: acceptance amongst
- Researchers
- Research administrators
- University governing boards
- Funders

Effectiveness of any new assessment scheme

=

Quality → More **valid indicators** & assessment procedures that combat biases

x

Acceptance

First-mover / collective action problem → CoARA is the best shot so far …

Transparency of requirements & **reproducibility** of indicators: No more proprietary black box algorithms

Efficiency in the hiring committee:
Can it handle 100+ applicants?

Cultural change: acceptance amongst
- Researchers
- Research administrators
- University governing boards
- Funders

The proposal, in a nutshell

inspired by …

DORA

RECOGNITION & REWARDS
ROOM FOR EVERYONE'S TALENT

CoARA

LERU

The Hong Kong Principles for assessing researchers: Fostering research integrity

David Moher[1,2]*, Lex Bouter[3,4], Sabine Kleinert[5], Paul Glasziou[6], Mai Har Sham[7], Virginia Barbour[8], Anne-Marie Coriat[9], Nicole Foeger[10], Ulrich Dirnagl[11]

DFG Deutsche Forschungsgemeinschaft

Research Funding | Funded Projects | DFG in Profile

Home › Funding › Calls for Proposals - Information for Researchers › 2022 › Package of Measures to Support a Shift in th

Information for Researchers No. 61 | 1 September 2022

Package of Measures to Support a Shift in the Culture of Research Assessment

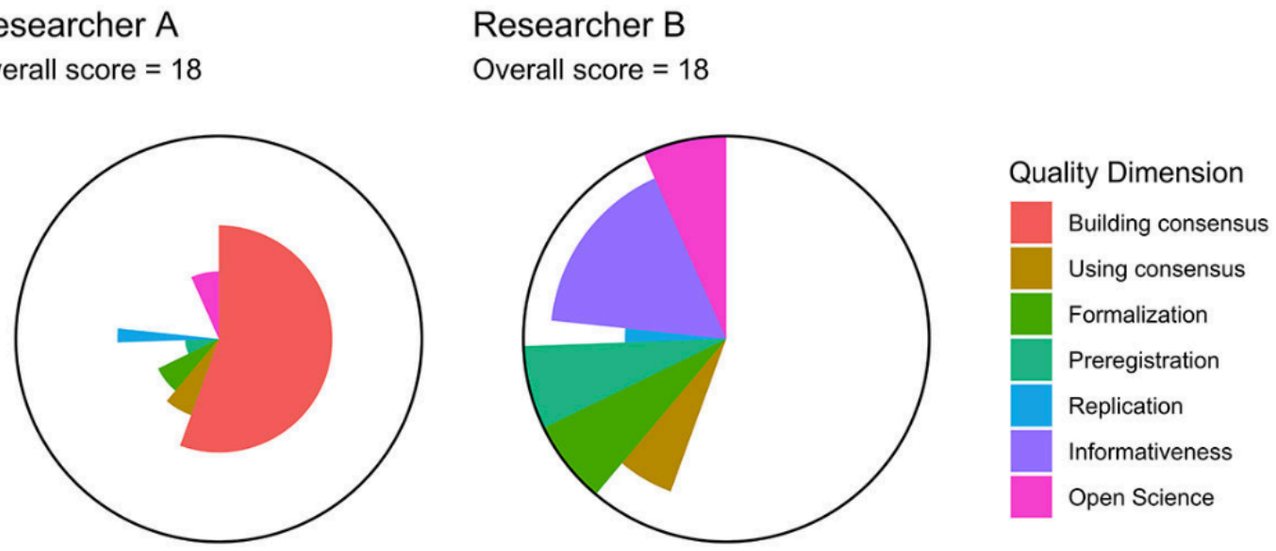
1. Expand the range of academic contributions

Types of academic contributions:

1. Research

Figure 1

Visualizing the Research Quality Profiles of two Researchers (A and B) who Promote Good Science in Different Ways, Through Their Respective Activities



Note. The width of each wedge is proportional to the maximum number of points that may be obtained in each category.

Figure from [Leising et al. (2022)]

2. Move from authorship to contributorship

Types of academic contributions:

1. Research

2. Teaching

3. Leadership
(e.g., mentoring, management and organizational skills, strategic thinking)

4. Service to the academic institution/ field

5. Societal impact
(e.g., science communication/ citizenship)

Contributor roles

Alternative: **Reorder** authors alphabetically in CV; only rely on CRediT

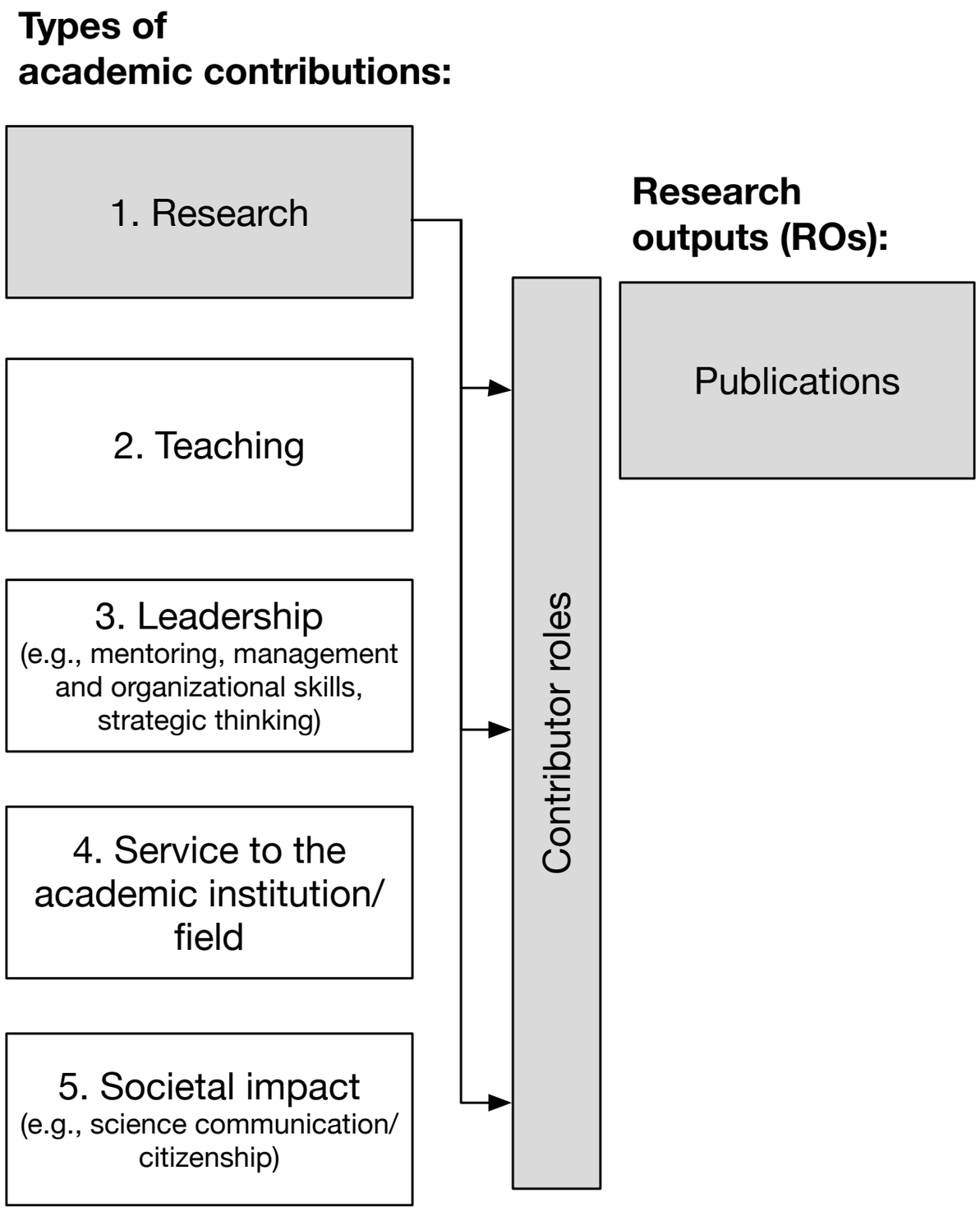
Reference	**CRediT** **L = lead,** **E = equal** **S = supporting**
Schönbrodt, F. D. & Wagenmakers, E.-J. (2018). Bayes Factor Design Analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128-142. doi:10.3758/s13423-017-1230-y	Conceptualization (L) Formal analysis (L) Methodology (L) Software (L) Writing – original draft (L)
Zygar, C., Hagemeyer, B., Pusch, S., & **Schönbrodt, F. D.** (2018). From motive dispositions to states to outcomes: An intensive experience sampling study on communal motivational dynamics in couples. *European Journal of Personality, 32*, 306–324. doi:10.1002/per.2145	Conceptualization (E) Data curation (S) Formal analysis (S) Funding acq (L) Investigation (S) Supervision (L) Writing – original draft (S)
Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., … **Schönbrodt, F. D.**, ..., & Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.	Writing – Review and editing (S)

JIF: 24!

Citations: 2239!

My contribution: Nearly zero.

3. More than publications: Data sets and software as fully-fledged contributions

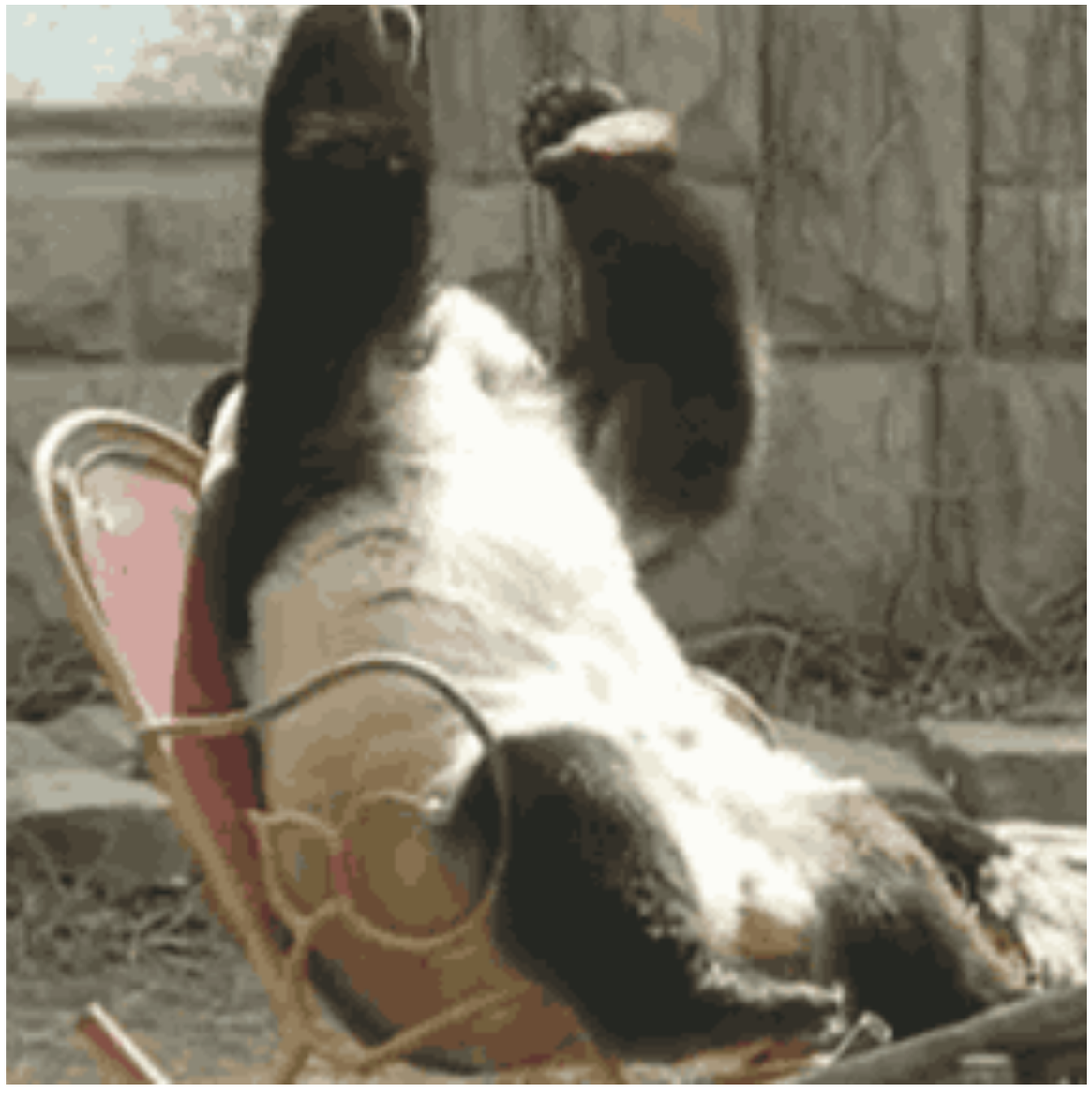


Quality over quantity

J.J. at the English language

4. Valid indicators for measuring quality (methodological rigor), impact, and quantity

In this work, we use the number of citations as a proxy for quality,



Quality over quantity

- „Quality“ is multidimensional:
  - basic aspects (methodological rigor)
  - elusive and complex aspects (innovation, creativity, ingenuity)
- Rigor - as one part of quality - can be measured (quite) objectively:
  Whether research has been **skillfully executed according to standards of good scientific practice within the field**.
- Quality cannot be *reduced* to rigor!
  - Rigor is *not a sufficient* condition for high-quality research –
    but it can be seen as a *necessary* condition for valid knowledge.

A quality/quantity tradeoff?

- QQ trade-off – a negative correlation between rigor and quantity –on the *within-person* level: Perceived extra effort & opportunity costs of open science practices (e.g., Houtkoop et al., 2018)
- Independent *between-person* effect: Some researchers are more capable of producing research outputs of high quality *and* higher quantity than others.
  - main target of assessment procedures
  - some level of quantitative productivity is certainly necessary for a researcher to be regarded as successful.
  - Once a good quality of products has been established, we may start counting them.
- However, the current practice of selecting competitors mainly via indicators of pure quantity, combined with a widespread *lack* of proper quality controls, sets an incentive for everybody to invest into the quantity, rather than the quality, of their own research.

4. Valid indicators for measuring quality (methodological rigor), impact, and quantity

Research outputs (ROs):	**Evaluation dimension: Quality / Rigor**	**Impact**	**Quantity**
Publications	- Registered report - Analysis script provided - Open material - Independently verified reproducibility - Formal modeling - Manipulation checks - Follows reporting guidelines - …	- Citation count - Altmetrics - Societal impact - …	- Number of papers - …
Data sets	- FAIRness - Representativeness - Size - Uniqueness/effort of data collection - …	- Citation count - # of reuses from other authors - …	- Number of published data sets - …
Research software	- Independent review - Unit testing - Documentation - Technology Readiness level - …	- Citation count - Dependencies - Github stars - …	- Number of published software - Duration of active maintenance - % of applicants contribution to a product - …

4. Valid indicators for measuring quality (methodological rigor) of papers

Used in Phase 1 algorithm?	ID	Criterion	Paper				
no	1	**Paper number (1, 2, 3, ...)**	*please specify*				
no	1	**Paper title (do *not* provide the journal name)**	*please specify*				
no	2	**Year of publication**	*please specify*				
no	3	**DOI**	*please specify*				
yes (as a filter: only show points and calculate score for empirical paper)	4	**Paper type (check all that apply)**	Empirical paper	☐			
			Meta-Analysis	☐			
			(Mainly) Theoretical contribution / Review	☐			
			Simulation	☐			
no	5	**CRediT roles (check all that apply to your contribution), see here for further information**	**CRediT role specification:**	**Lead**	**Equal**	**Support**	**No Role**
			Conceptualization	○	○	○	◉
			Data curation	○	○	○	◉
			Formal analysis	○	○	○	◉
			Funding acquisition	○	○	○	◉
			Investigation	○	○	○	◉
			Methodology	○	○	○	◉
			Project administration	○	○	○	◉
			Resources	○	○	○	◉
			Software	○	○	○	◉
			Supervision	○	○	○	◉

(this is only a selection of indicators)

4. Valid indicators for measuring quality (methodological rigor) of papers

ID	Criterion	Paper 1	
9	**Correctness of computational results has been independently verified**	No	☐
		Not applicable	☐
		Yes [provide doi or URL to verification report]	☑
10	**Open reproducible scripts**	Not available	☐
		Not applicable [provide explanation]	☐
		Yes [provide doi or URL]	☑
11	↳ **FAIR format (conditional: only when (10) is "Yes")**	Timestamped repository	☑
		Version control	☑
		Reproducible manuscripts (e.g. with R Markdown)	☑
		Reproducible software environments (e.g. conda environment, renv environment in R)	☑

(this is only a selection of indicators)

4. Valid indicators for measuring quality (methodological rigor) of papers

ID	Criterion	Paper	
19	**Citation count from Google Scholar**	*please specify*	
20	**Merit / impact statement (narrative, max. 150 words)**	*please specify (please also provide supporting documents or links)*	
21	**Flag: Does this contribution fit well into the assessment scheme?**	Yes	☐
		No, it should be processed manually [provide explanation]	☐

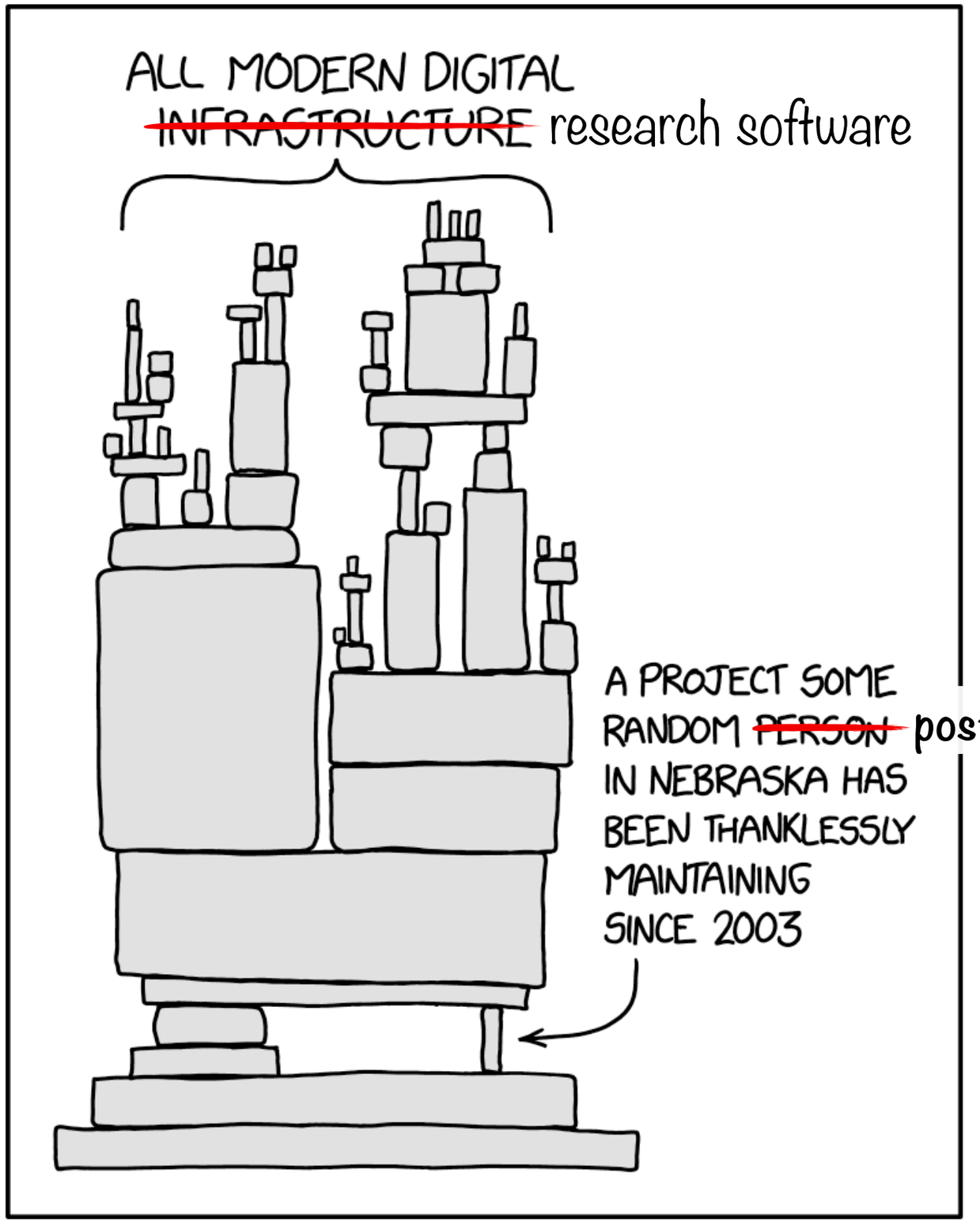
(this is only a selection of indicators)

4. Valid indicators for measuring quality (methodological rigor) of open data sets

ID	Criterion	Data Set 1	
4	**DataCRediT roles (check all that apply to your contribution), see here for further information**	Collection	☐
		Validation	☐
		Curation	☐
		Software	☐
		Publication	☐
		Supervision	☐
5	**Data type(s) (check all that apply)**	Questionnaire	☐
		Behavioral	☐
		Physiological / Biological	☐
		Other:	☐
6	**Study mode (check all that apply)**	Online	☐
		Laboratory	☐
		Field study	☐
		Experience sampling	☐
		Other:	☐
7	↳ **FAIR format (see here for further information)**	No	☐
		To some extent	☐
		Completely FAIR	☑

(this is only a selection of indicators)

4. Valid indicators for measuring quality (methodological rigor) of research software



https://xkcd.com/2347/

4. Valid indicators for measuring quality (methodological rigor) of research software

ID	Criterion		Comment
7	**Contributor roles and involvement**	DA-3 CD-3 MS-3	What has the applicant contributed? For each of the 3 roles: - design and architecture (DA) - coding and debugging (CD) - maintenance and support (MS) Specify if you are: 0: not involved 1: an occasional contributor 2: a regular contributor 3: a main contributor Example: DA-2, CD-3, MS-1
8	**License**	GPLv3	Is the software open source?

(this is only a selection of indicators)

4. Valid indicators for measuring quality (methodological rigor) of research software

16	Years of active involvement	How many years have you significantly contributed to the software project, in either role?
17	Reusability indicator	**Levels of the reusability indicator**: **R1 (0.25 points)**: Single scripts, loose documentation, no long-term maintenance. *Prototype*: A collection of reusable *Python* scripts on OSF. **R2 (2 points)**: Well-developed and tested software, fairly extensive documentation. Some attention to usability and user feedback. Not necessarily regularly updated. *Prototype:* A small *CRAN* package with no more active development (just maintenance). **R3 (4 points)**: Major software project, strong attention to functionality and usability, extensive documentation, systematic bug chasing and unit testing, external quality control (e.g. by uploading to a registry like *npm*, *CRAN* or *juliapackages*). Regularly updated. *Prototype*: Well received and actively maintained Julia package. **R4 (8 points)**: Critical infrastructure software. Hundreds of research projects use or depend on the software (+ all criteria of R3). *Prototype*: *lavaan* or *brms* package for R.

(this is only a selection of indicators)

With the suggested reward points, a „critical infrastructure software“ can count as much as 5 good research papers.

Effectiveness of any new assessment scheme

=

Quality → More **valid indicators** & assessment procedures that combat biases

x

Acceptance

First-mover / collective action problem ➙
CoARA is the best shot so far …

Transparency of requirements & **reproducibility** of indicators: No more proprietary black box algorithms

Efficiency in the hiring committee:
Can it handle 100+ applicants?

Cultural change: acceptance amongst
- Researchers
- Research administrators
- University governing boards

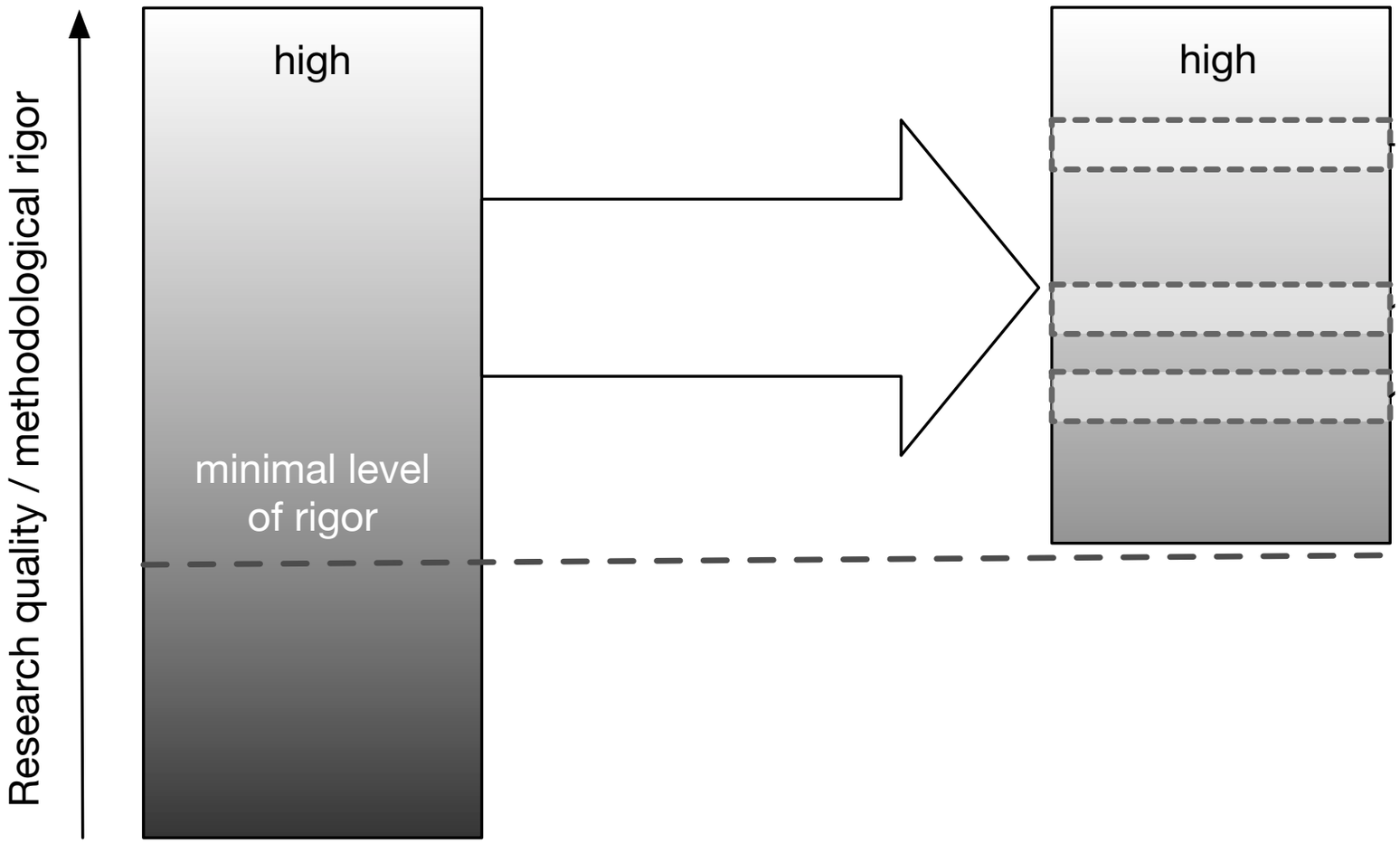
5. A two-phase hiring process

Phase 1:
Negative selection with focus on *efficiency*: algorithm/indicator assisted

Longlist
(applicants who formally fit to the job description)

→

Shortlist
(candidates generally qualified for the job)



5. A two-phase hiring process: Efficiency

- Candidates are expected to enter the indicator values themselves for, e.g., their 10 best publications, up to 5 data sets and up to 5 pieces of software.
- Out of these 20 research outputs, the candidates nominate their 3 best works (for in-depth evaluation in phase 2)
- Someday, these information could go into a database and need only to be entered once.
- Research assistants can do most of data collection.
  - Pilot study: ICC(1,1) = .81
- Self-reports should be checked on a random basis (longlist) or routinely for all candidates on the shortlist.
- Phase 1 can be algorithmically assisted (but do not bypass human judgement completely!)
  - Templates for aggregating and visualizing the provided information → multidimensional profiles

5. A two-phase hiring process

Phase 1:
Negative selection with focus on *efficiency*: algorithm/indicator assisted

Phase 2:
Positive selection with focus on *content*: in-depth qualitative evaluation and peer-review in committee

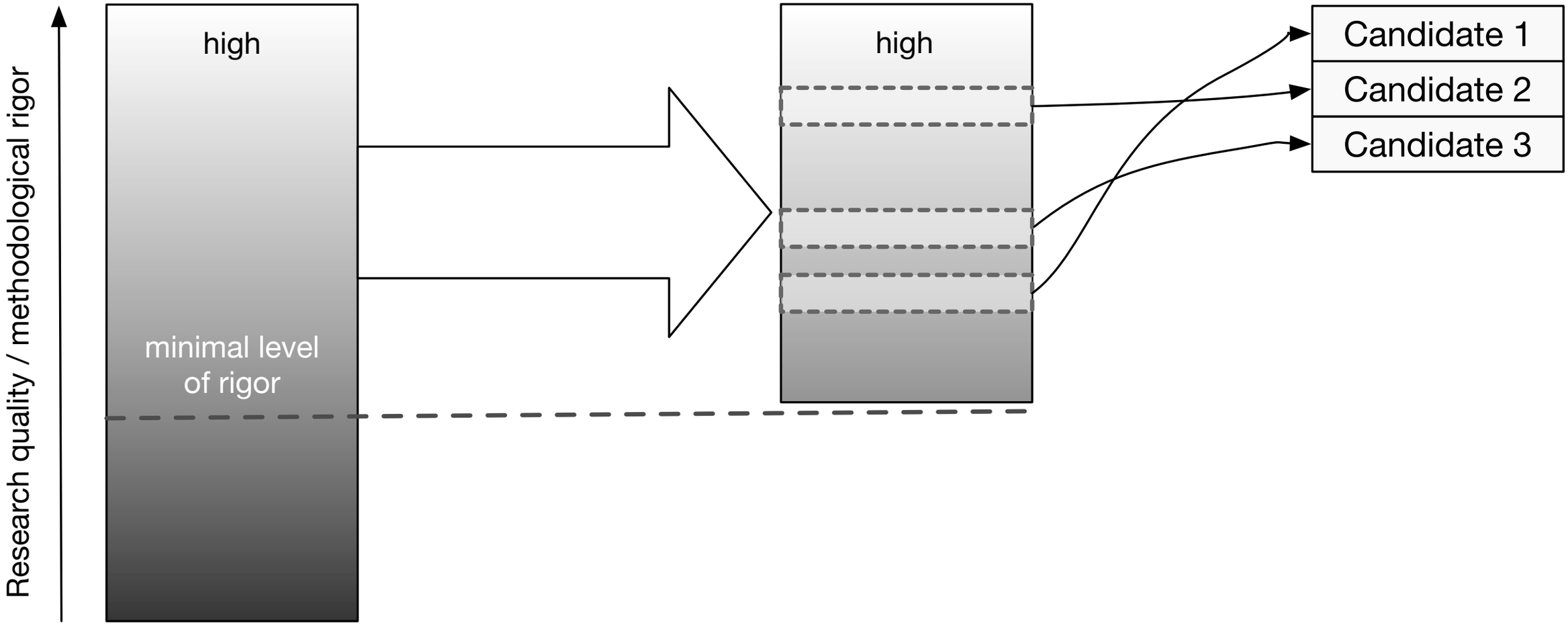
Longlist
(applicants who formally fit to the job description)

→

Shortlist
(candidates generally qualified for the job)

→

Final list with ranked candidates



Effectiveness of any new assessment scheme

=

Quality ⟶ More **valid indicators** & assessment procedures that combat biases

x

Acceptance

First-mover / collective action problem ➙
CoARA is the best shot so far …

Transparency of requirements & **reproducibility** of indicators: No more proprietary black box algorithms

Efficiency in the hiring committee:
Can it handle 100+ applicants?

Cultural change: acceptance amongst
- Researchers
- Research administrators
- University governing boards

The (social) process

DGPs Deutsche Gesellschaft für Psychologie

2022
- Expert task group commissioned by the German Psychological Society

09/2022
- First draft of recommendations discussed at national conference

12/2022
- **Draft published as target paper** in the diamond open access journal *Meta-Psychology*

03/2023
- **Community feedback** by published commentaries

Commentary: 'Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology'

Andreas M. Brandmaier[1,2,3], Maximilian Ernst[2,4], &

Responsible Research is also concerned with generalizability:

Recognizing efforts to reflect upon and increase generalizability in hiring and promotion decisions in psychology

Roman Stengelin[1,2]; Manuel Bohn[1]; Alejandro Sánchez-Amaro[1]; Daniel B.M. Haun[1,3,4]; Maleen Thiele[1]; Moritz M. Daum[5,6]; Elisa Felsche[1]; Frankie T.K. Fong[1,7]; Anja Gampe[8]; Marta Giner Torréns[9]; Sebastian Grueneisen[3,10]; David J.K. Hardecker[1,3]; Lisa Horn[11]; Karri Neldner[1,4]; Sarah Pope-Caldwell[1]; Nils Schuhmacher[9]

Interdisciplinary Value

Veli-Matti Karhulahti*
University of Jyväskylä

* Corresponding author. Email vmkarhwu@jyu.fi

Abstract. This is a commentary on the following two studies, which serve as target articles in the *Psychology* special issue "Responsible Research Assessment: Implementing DORA for hiring promotion in psychology".

Commentary: "Responsible Research Assessment: Implementing DORA for hiring and promotion in psychology"

Alejandro Sandoval-Lentisco[1]

[1] *Dept. Basic Psychology and Methodology, University of Murcia, Murcia, Spain*

Responsible Research Assessment Should Prioritize Theory Development and Testing Over Ticking Open Science Boxes

Authors:

Hannah Dames[1], Philipp Musfeld[1], Vencislav Popov[1], Klaus Oberauer[1], Gidon T. Frischkorn[1]

15 commentaries

The (social) process

DGPs Deutsche Gesellschaft für Psychologie

2022
- Expert task group commissioned by the German Psychological Society

09/2022
- First draft of recommendations discussed at national conference

12/2022
- **Draft published as target paper** in the diamond open access journal *Meta-Psychology*

03/2023
- **Community feedback** by published commentaries

2023
- **Revised version** of recommendations; ready-to-use templates ➙ hopefully official recommendation by DGPs
- Application and evaluation: Rapid prototyping process with continuous evaluation and refinement
  - Assess interrater-reliability & validity of indicators
  - Explore automatic coding
    (e.g. ScreenIT, DataSeer, Rigor and Transparency Index)
  - Feed insights back to CoARA

2024

Discussion

- What are potential negative side-effects?
- Goodhart's law: How could you hack the new system?
- Barriers for implementation: What would the chair of your next hiring committee say when you propose to switch to the new system?

References & Resources

- Schönbrodt, F. D., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., … Leising, D. (2022, November 25). **Responsible Research Assessment I: Implementing DORA for hiring and promotion in psychology**. https://doi.org/10.31234/osf.io/rgh5b

- Gärtner, A., Leising, D., & Schönbrodt, F. D. (2022, November 25). **Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology**. https://doi.org/10.31234/osf.io/5yexm

- **Call for commentaries** at Meta-Psychology (deadline: 2023-03-31): https://open.lnu.se/index.php/metapsychology/announcement/view/5

- Gärtner, A., Leising, D., & Schönbrodt, F. D. (2023, March 3). Empfehlungen zur Berücksichtigung von wissenschaftlicher Leistung bei Berufungsverfahren in der Psychologie. https://doi.org/10.31234/osf.io/3yjz7

- Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – **how quality may be rewarded more in research evaluation**. Personality Science, 3, e6029. https://doi.org/10.5964/ps.6029

- OSF project with **templates of grading sheets**: https://osf.io/4wynr/wiki/home/

Speicher

Subtitle

Why is the uptake so slow?

1. No idea how to do it better
2. Too much effort
3. Restricting the academic freedom of committees?
4. Social dilemma: First movers have a disadvantage
5. Committee members maybe excelled on the old metrics (but not necessarily on the new ones?)
6. A sudden change in assessment criteria is unfair (after all, we spent years optimizing the old ones)

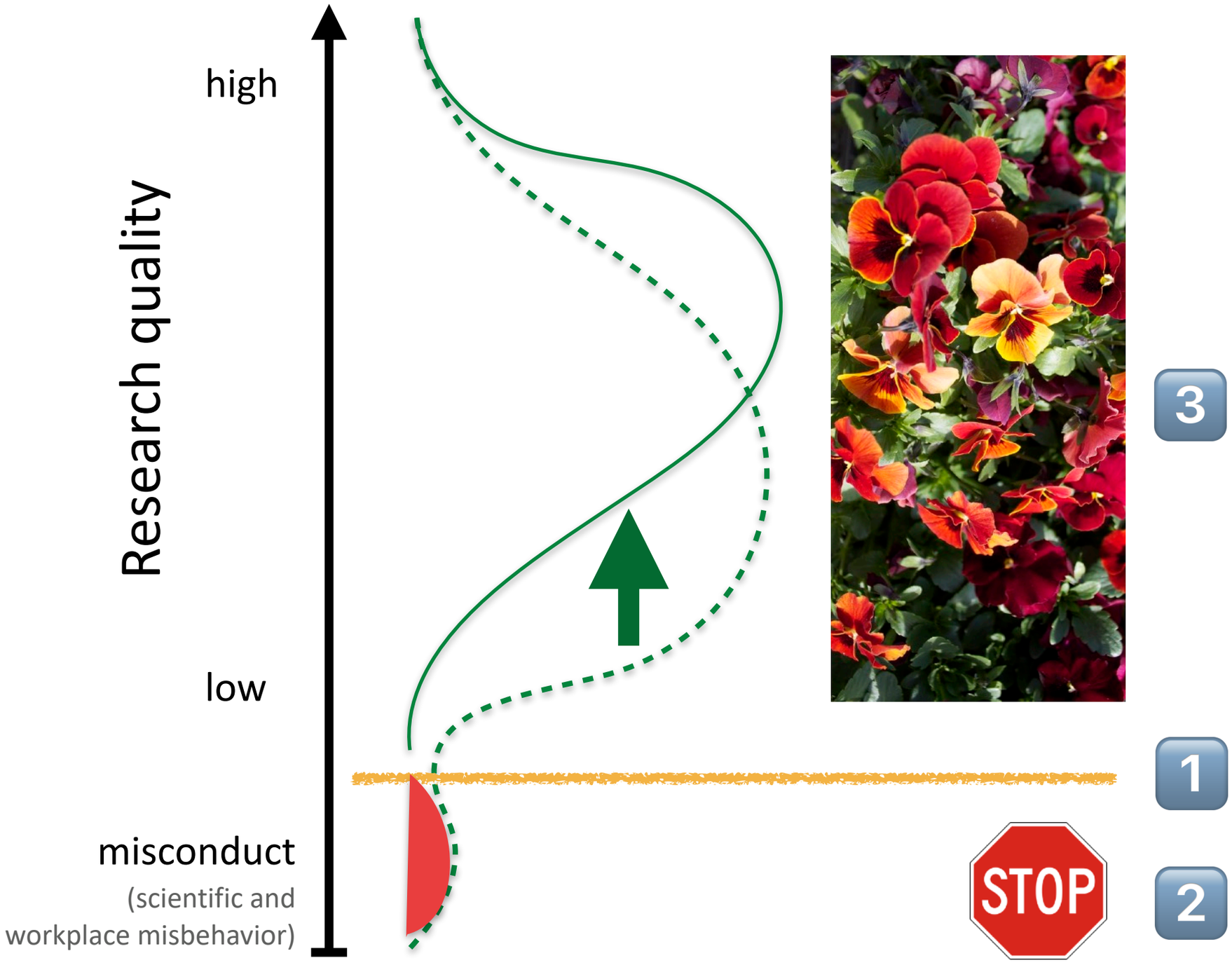
Scientific progress I

- When the goal is scientific progress, defined as achieving valid and credible knowledge, it is important to differentiate *progress* and *quality*:
- "**Quality** is primarily an activity-oriented concept, concerning the skill and competence in the performance of some task.
- **Progress** is a result-oriented concept, concerning the success of a product relative to some goal.
- All acceptable work in science has to fulfill certain standards of quality. But it seems that there are no necessary connections between quality and progress in science. Sometimes very well-qualified research projects fail to produce important new results, while less competent but more lucky works lead to success.
- *Nevertheless, the skillful use of the methods of science will make progress highly probable. Hence, the best practical strategy in promoting scientific progress is to support high-quality research.*" (Niiniluoto, 2019, p. 6).

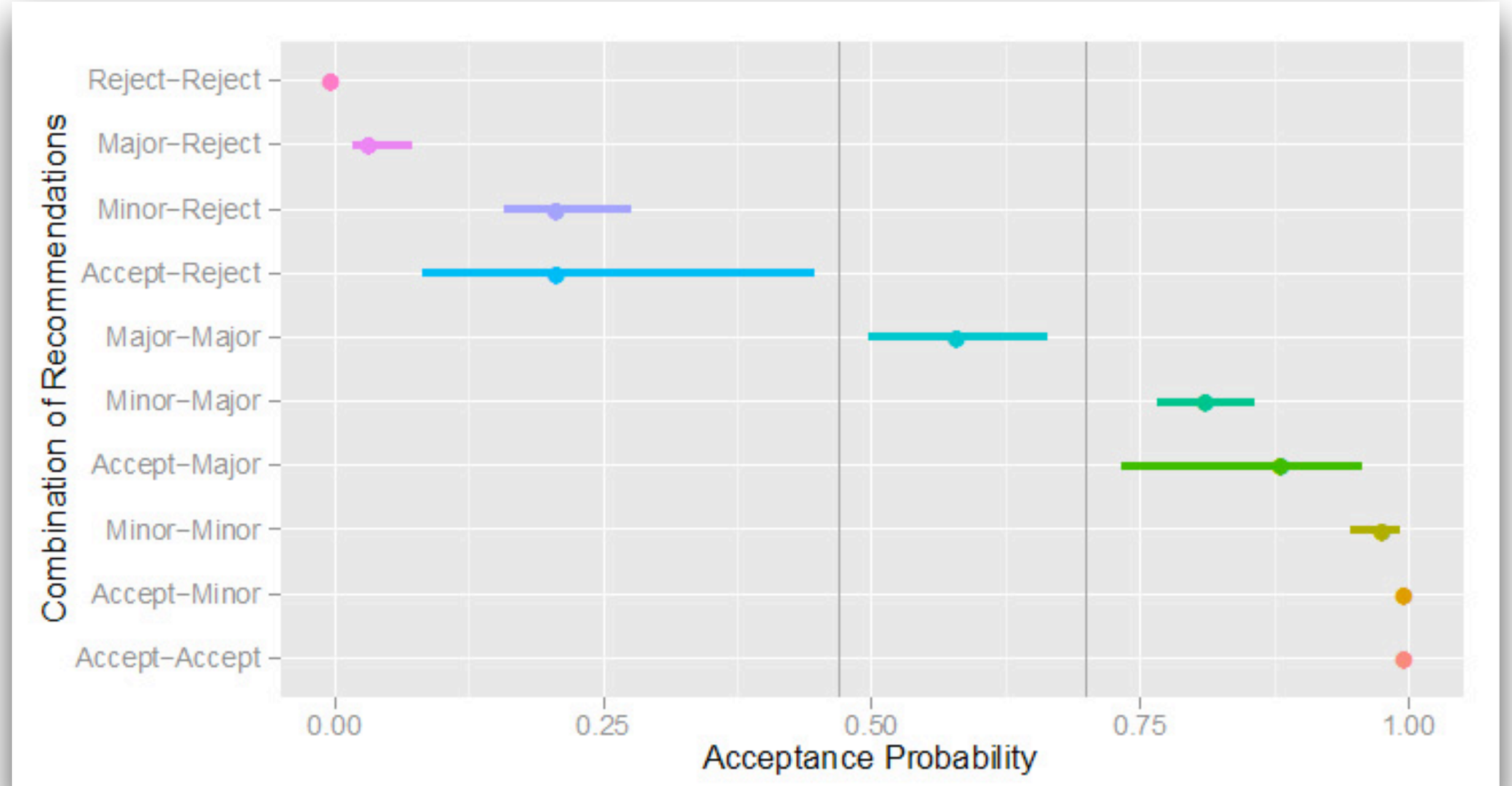
Scientific progress II

- Assumption 1: We will never be able to predict what research will be excellent, useful, or impactful (in the real world).
- Assumption 2: We know quite well what bad science is.
- Solution to foster scientific progress:
  - Weed out bad science
  - Support researchers to achieve high standards of methodological rigor
  - (See next slide)

1. Scientific fields should debate and find a consensus about the basic level of necessary good research practices („craftsmanship“).
2. These should be required, controlled and enforced by universities, funders, journals, supervisors.
3. Those who comply to this minimal standard should be free to thrive, with as few regulations and bureaucratic compliance as possible.



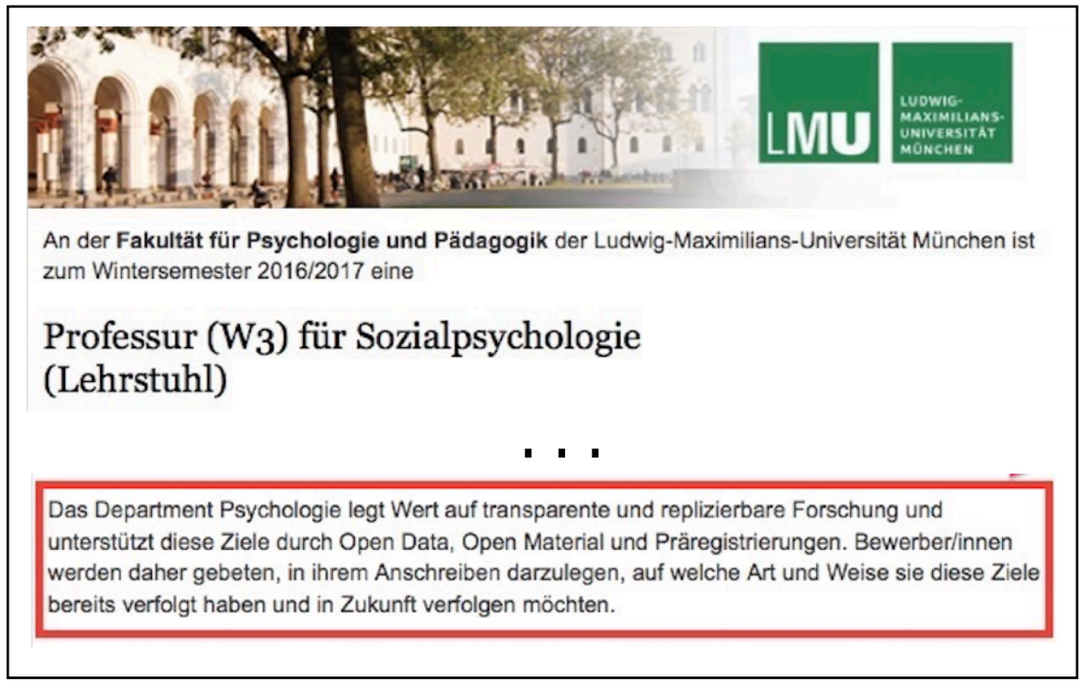
„But the reviewers do not decide about rejection and acceptance - the editor weighs several sources of information to reach an informed decision.“
➞ in the ideal case, yes. But in general, the decision is closely related to the reviewers assessment:



https://www.elsevier.com/connect/is-peer-review-just-a-crapshoot

Changing the incentive structure: Professorship hiring committees

Hiring committees: Make „open science“ a desirable or essential job characteristic



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

An der **Fakultät für Psychologie und Pädagogik** der Ludwig-Maximilians-Universität München ist zum Wintersemester 2016/2017 eine

Professur (W3) für Sozialpsychologie
(Lehrstuhl)

. . .

Das Department Psychologie legt Wert auf transparente und replizierbare Forschung und unterstützt diese Ziele durch Open Data, Open Material und Präregistrierungen. Bewerber/innen werden daher gebeten, in ihrem Anschreiben darzulegen, auf welche Art und Weise sie diese Ziele bereits verfolgt haben und in Zukunft verfolgen möchten.

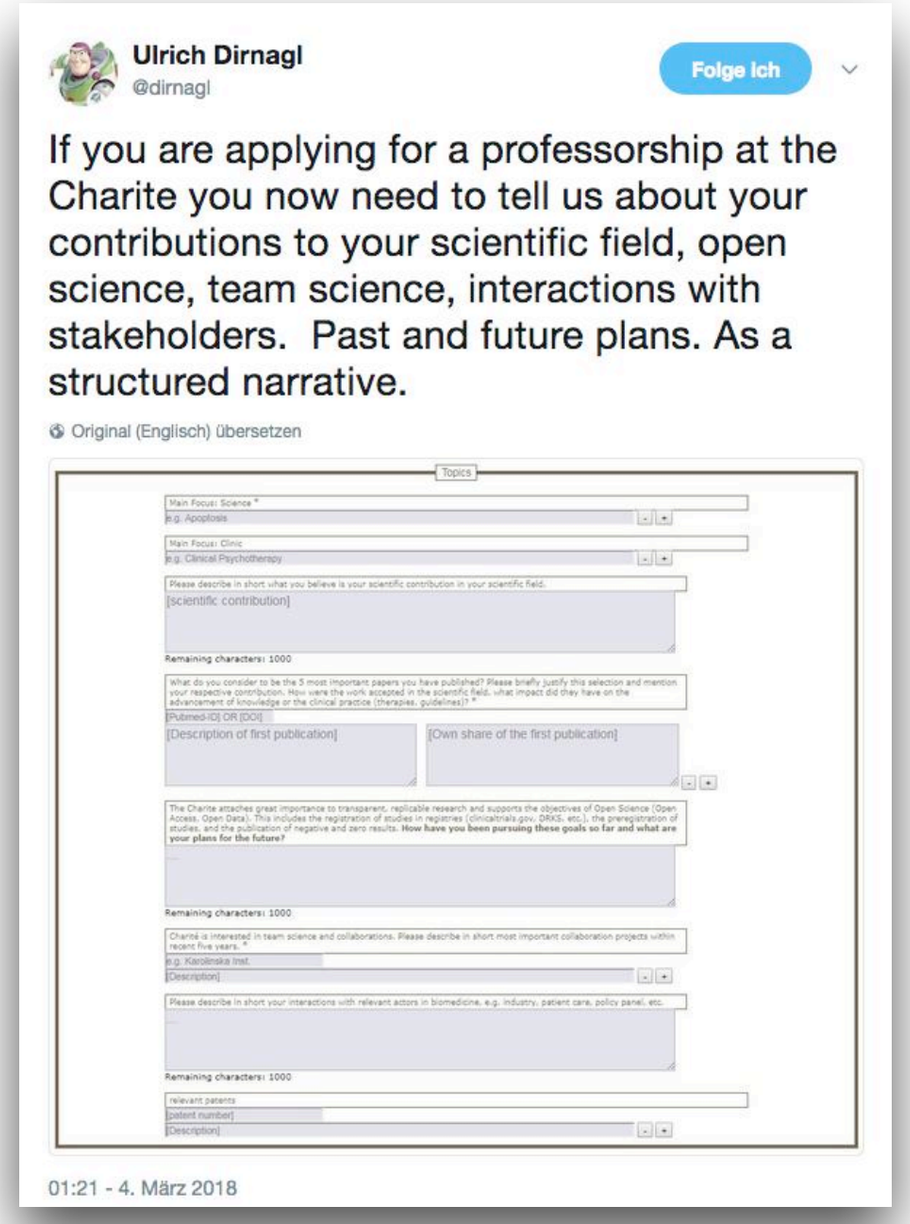
Since 2015: All professorship job descriptions use this requirement

Full Professor (W3)
of Social Psychology

to be filled as soon as possible.

. . .

The Department of Psychology aims for transparent and reproducible research (including Open Data, Open Materials, and Preregistrations). Applicants are asked to illustrate how they have pursued these goals in the past and/or how they plan to do so in the future.



Ulrich Dirnagl
@dirnagl

Folge ich

If you are applying for a professorship at the Charite you now need to tell us about your contributions to your scientific field, open science, team science, interactions with stakeholders. Past and future plans. As a structured narrative.

Original (Englisch) übersetzen

Topics

Main Focus: Science *
e.g. Apoptosis

Main Focus: Clinic
e.g. Clinical Psychotherapy

Please describe in short what you believe is your scientific contribution in your scientific field.
[scientific contribution]
Remaining characters: 1000

What do you consider to be the 5 most important papers you have published? Please briefly justify this selection and mention your respective contribution. How were the work accepted in the scientific field, what impact did they have on the advancement of knowledge or the clinical practice (therapies, guidelines)? *
[Pubmed-ID] OR [DOI]
[Description of first publication]
[Own share of the first publication]

The Charite attaches great importance to transparent, replicable research and supports the objectives of Open Science (Open Access, Open Data). This includes the registration of studies in registries (clinicaltrials.gov, DRKS, etc.), the preregistration of studies, and the publication of negative and zero results. **How have you been pursuing these goals so far and what are your plans for the future?**
Remaining characters: 1000

Charité is interested in team science and collaborations. Please describe in short most important collaboration projects within recent five years. *
e.g. Karolinska Inst.
[Description]

Please describe in short your interactions with relevant actors in biomedicine, e.g. industry, patient care, policy panel, etc.
Remaining characters: 1000

relevant patents
[patent number]
[Description]

01:21 - 4. März 2018

See more such prof job ads at:

https://docs.google.com/document/d/1ty43Syw0Flkh8ncjW8MZArIkvYe8hLwwhLIlwbtSk_Y/edit?usp=drive_web&ouid=108982640291853577145

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR PSYCHOLOGIE UND PÄDAGOGIK

Google™ benutzerdefinierte Su | www.lmu.de | Sitemap

FAKULTÄT
STUDIUM
FORSCHUNG
PSYCHOLOGIE
Studium und Lehre
Forschung
Open-Science-Committee
About our OSC
Recognizing Open Research Practices in Our Hiring Policy
Workshops and Talks
Lehr- und Forschungseinheiten
Ambulanzen und Testlab

drucken

Recognizing Open Research Practices in Our Hiring Policy

In December 2015, the Department Psychology of the LMU Munich added a paragraph to a professorship announcement which emphasized the department's commitment to responsible research and asked applicants to write a short statement about their open science practices:

"Our department embraces the values of open science and strives for replicable and reproducible research. For this goal we support transparent research with open data, open materials, and study pre-registration. Candidates are asked to describe in what way they already pursued and plan to pursue these goals."

Since then, all further professorship job advertisements of our department had this requirement.

In May 2018, the department's steering committee unanimously voted for an explicit policy to always include this (or a similar) statement to all future professorship job advertisements. It is the task of the appointment committee to value the existing open science activities as well as future commitments of applicants appropriately. By including this statement, our department aims to communicate core values of good scientific practice and to attract excellent researchers who aim for transparent and credible research.

http://www.fak11.lmu.de/dep_psychologie/osc/open-science-hiring-policy/index.html

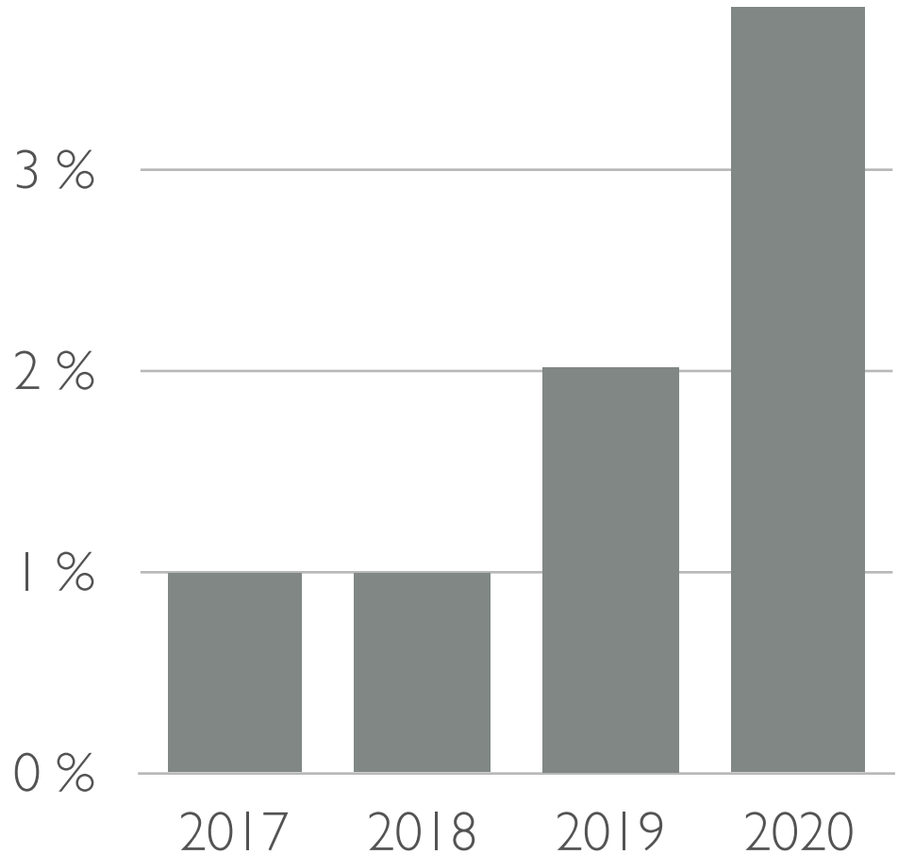
Change of incentive structures: Hiring policy

OSC
LMU Open Science Center

Analysis of job offers in the field of psychology:

- 1626 job ads (1484 in German, 142 in English); entire database of from February 2017 to December 2020
- Keyword search for *open science*, *reproduc**, *replication*, *research transparency*, etc.
- Out of 376 advertising institutions, 20 mentioned replicability and transparency at least once.
- Across all analyzed years, 2.2% (n=36) of job offers mentioned replicability and transparency as desired or essential job criteria.

academics
Aus dem Hause DIE ZEIT Forschung & Lehre



Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., …, Schönbrodt, F. D., Vazire, S. (2021, February 9). *Replicability, Robustness, and Reproducibility in Psychological Science.* https://doi.org/10.31234/osf.io/ksfvq