

Responsible Research Assessment

Evaluating research quality beyond *h*-index and JIF
with the RESQUE framework
(RESearch QUality Evaluation)

Felix Schönbrodt, Anne Gärtner, Daniel Leising

*Alp Kaan Aksu, Maximilian Ernst, Maximilian Frank, Nele Freyer, Mario Gollwitzer, Malika Ihle,
Jens Lange, Dorothee Mischkowski, Philipp Musfeld, Aaron Peikert, Le Vy Phan, Manfred Schmitt,
Anne M. Scheel, Anna-Lena Schubert, Ulf Steinberg*

Eine typische Berufungskommission



Eine Karikatur.

(?)

Eine typische BK ...



- Man erhält unstrukturierte PDFs, aus denen man mühsam (und fehleranfällig) die relevanten Informationen heraussucht

Eine typische BK ...

- Ein Sekretariat erstellt eine xlsx-Datei mit Bewerberübersicht, mit den „üblichen“ Indikatoren.
 - Findet manche Bewerber nicht im Web of Science, nimmt dann Google Scholar
 - Unklar, ob Drittmittel nur als „PI“ gelten, ist oft auch gar nicht angegeben.

Name	Vorname	Anrede	Habilitation bzw. equivalente Leistung	Publikationen (Peer- Review Artikel)	Erstautorenschaften	Jahre der Publikationsstätigkeit	Publikationen/Jahr	Drittmittelprojekte (> 20.000 Euro)	Summe Drittmittel	Forschungsansatz (gewünscht: fMRT, EEG, tMS)	Anzahl Zitationen (Google Scholar)	h-Index
Mustermann	Max	Dr.	2019	31	11	8	3,9	4	640.000	fMRT, EEG	8.239	22
Musterfrau	Eva	Prof. Dr.	2015	13	4	13	1,0	2	1.100.000	fMRT, EEG	10.265	13
Einstein	Albert	Prof. Dr.	1895	8	8	7	0,5	0	0	(nicht empirisch)	172.073	8

Eine typische BK ...

- Sehr wenige Personen machen relativ subjektiv die erste Selektion (um die Kommission nicht zu überfordern)
- In der **Kommissionssitzung** ...
 - ... ist unklar wie Elemente aus der Ausschreibung gemessen/bewertet werden sollen (vgl. „Open-Science-Statement“)
 - ... ist unklar, wie die verschiedenen Dimensionen gewichtet werden sollen
 - ... wird von der Shortlist jeweils 1 Paper von 1 Kommissionsmitglied gelesen, das sein zusammenfassendes Urteil vorstellt
 - Eher nach Bauchgefühl - wirklich auskennen tut sich außer dem externen Mitglied niemand, da genau dieses Fach ja nicht besetzt ist.
 - ... wird in einer intuitiven Gewichtung eine Liste erstellt.



Desiderata eines Auswahlverfahrens

- **Standardisierung der verfügbaren Informationen:** Alle relevanten Informationen sollten von allen Kandidaten parallel vorliegen
 - dazu muss die Kommission **vor** der Ausschreibung festlegen, welche Bewertungsdimensionen wichtig sind, und diese einfordern
- **Effizienz:** Die Kommission soll sich auf die wichtigen Fragen konzentrieren können
 - Bei vielen Bewerbern: Effizienz und Validität in der ersten Selektion.
- Eine Aufbereitung der Bewerberinformationen, die dazu anregt, über **Forschungsqualität** (weniger -quantität) und **multidimensionale Profile** zu diskutieren
 - v.a. in der Shortlist-Phase eine inhaltliche Auseinandersetzung anregen.

Areas of research assessment

- Graduation (PhD, habilitation), hiring, tenure track
- Funding
- Awards, Rankings
- Performance-oriented payments and rewards

What is the goal of research assessment in academic hiring?

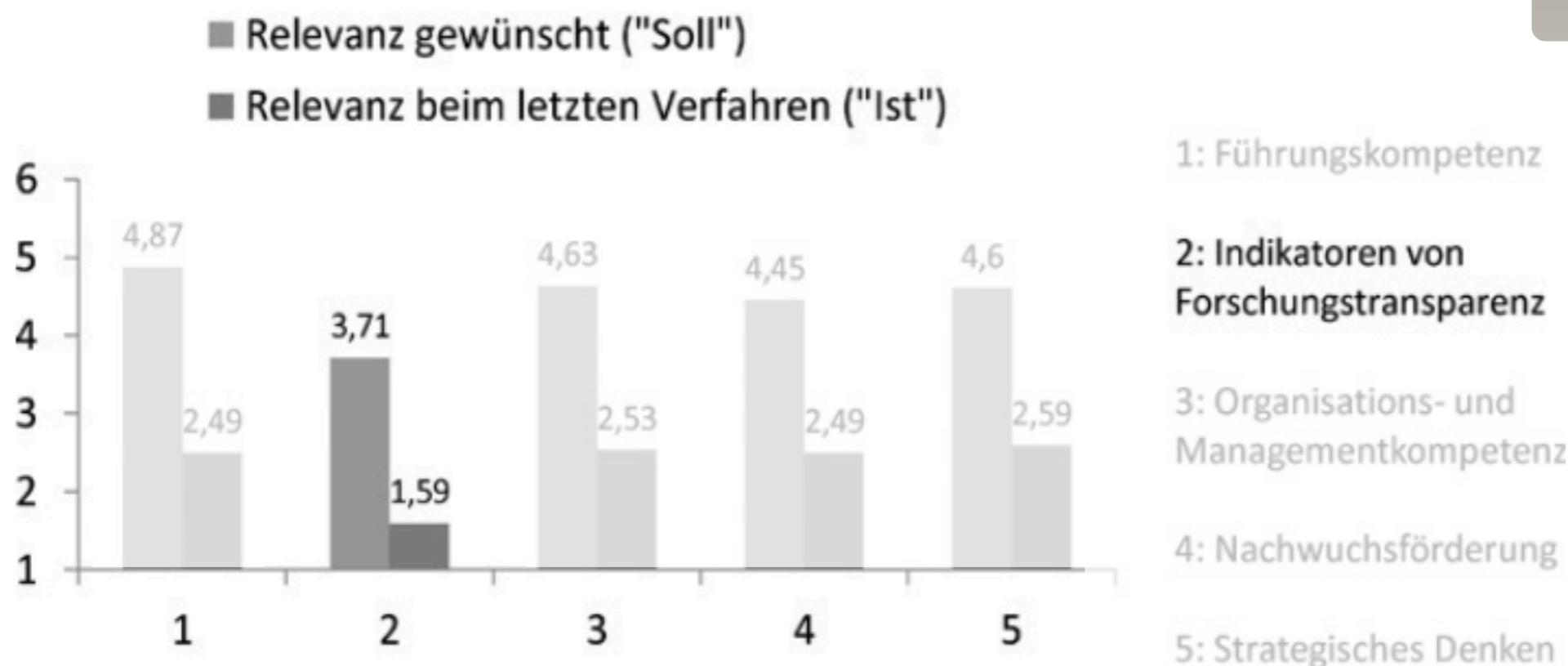
- To find a colleague who ...
 - brings a lot of third-party funding to the university?
 - boosts the university's THE ranking by having a lot of papers, which are also highly cited?
 - excels in teaching and mentoring?
 - contributes to scientific progress and credible knowledge?
 - is a nice and agreeable person who makes no trouble in the department?

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)

Quality, not quantity

Kriterien mit der größten Diskrepanz zwischen „Soll“ und „Ist“



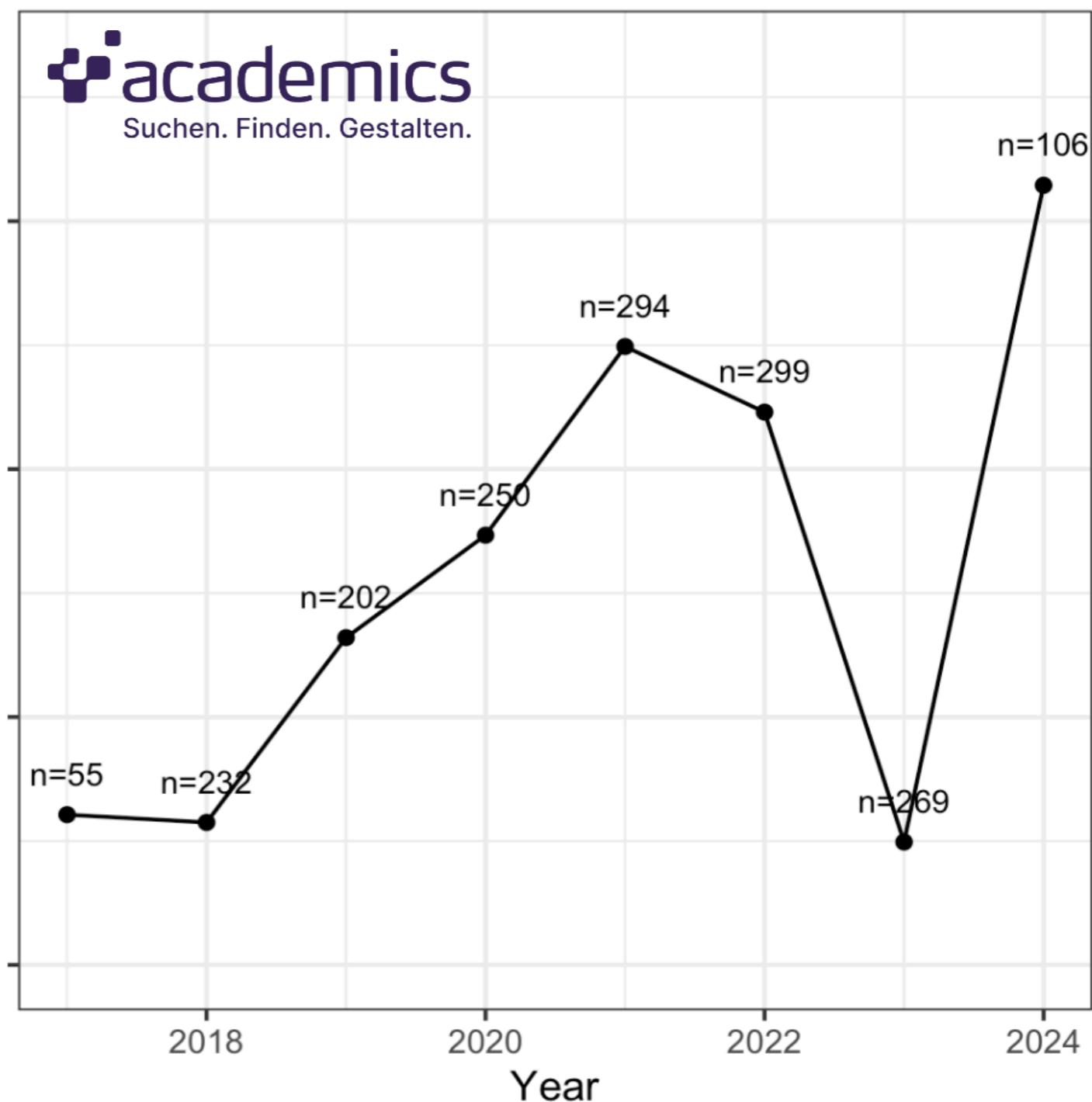
Indicators with the largest discrepancy between „desired“ and „actual“:
Researchers want to have indicators of research transparency in hiring committees!

Change of incentive structures: Hiring policy

Analysis of professorship job announcements in psychology:

- 1707 job ads; entire database of academics.de from Feb 2017 to June 2024
- Keyword search for *open science, reproduc**, *replication, research transparency*, etc.
- Out of 420 advertising institutions, 34 mentioned replicability and transparency at least once (8%) as desired or essential skill of a professor.

% of prof job ads mentioning open science



Responsible Research Assessment: A proposal for professorship hiring committees

Bewertungskriterien wissenschaftlicher Leistung



„1. Recognise the **diversity of contributions** to research.

- „practices that contribute to robustness, openness, transparency,
- „Inappropriate uses of journal- and publication-based metrics in research assessment should be abandoned. In particular, this means **moving away from using metrics like the Journal Impact Factor (JIF), Article Influence Score (AIS) and h-index** as proxies for quality and impact.“



„das ganze Spektrum wissenschaftlicher Publikationsformen gleichwertig in [...] Lebensläufen abzubilden“

„Dies können beispielsweise Artikel auf Preprint-Servern, Datensätze oder Softwarepakete sein.“

„Angaben zu quantitativen Metriken wie Impact-Faktoren und h-Indizes im Lebenslauf oder Antrag werden nicht benötigt und sollen in der Begutachtung keine Berücksichtigung finden.“

1. Expand the range of academic contributions

Types of academic contributions:

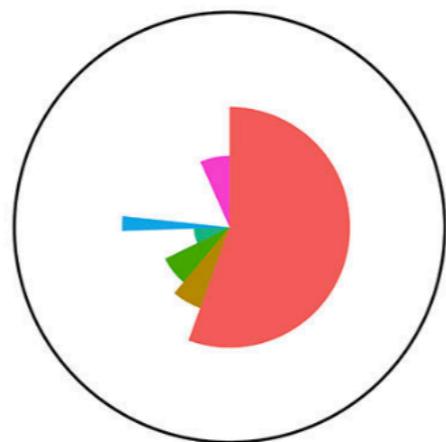
1. Research

Figure 1

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

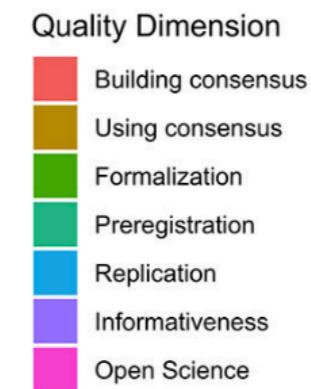
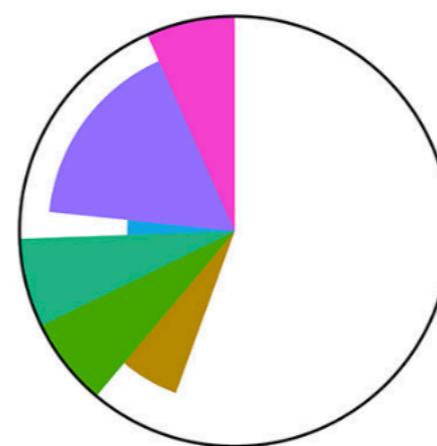
Researcher A

Overall score = 18



Researcher B

Overall score = 18



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)

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

Contributor roles

Reference

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

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

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.

CRediT

L = lead,
E = equal
S = supporting

Conceptualization (L)
Formal analysis (L)
Methodology (L)
Software (L)
Writing – original draft (L)

Conceptualization (E)
Data curation (S)
Formal analysis (S)
Funding acq (L)
Investigation (S)
Supervision (L)
Writing – original draft (S)

Writing – Review and editing (S)

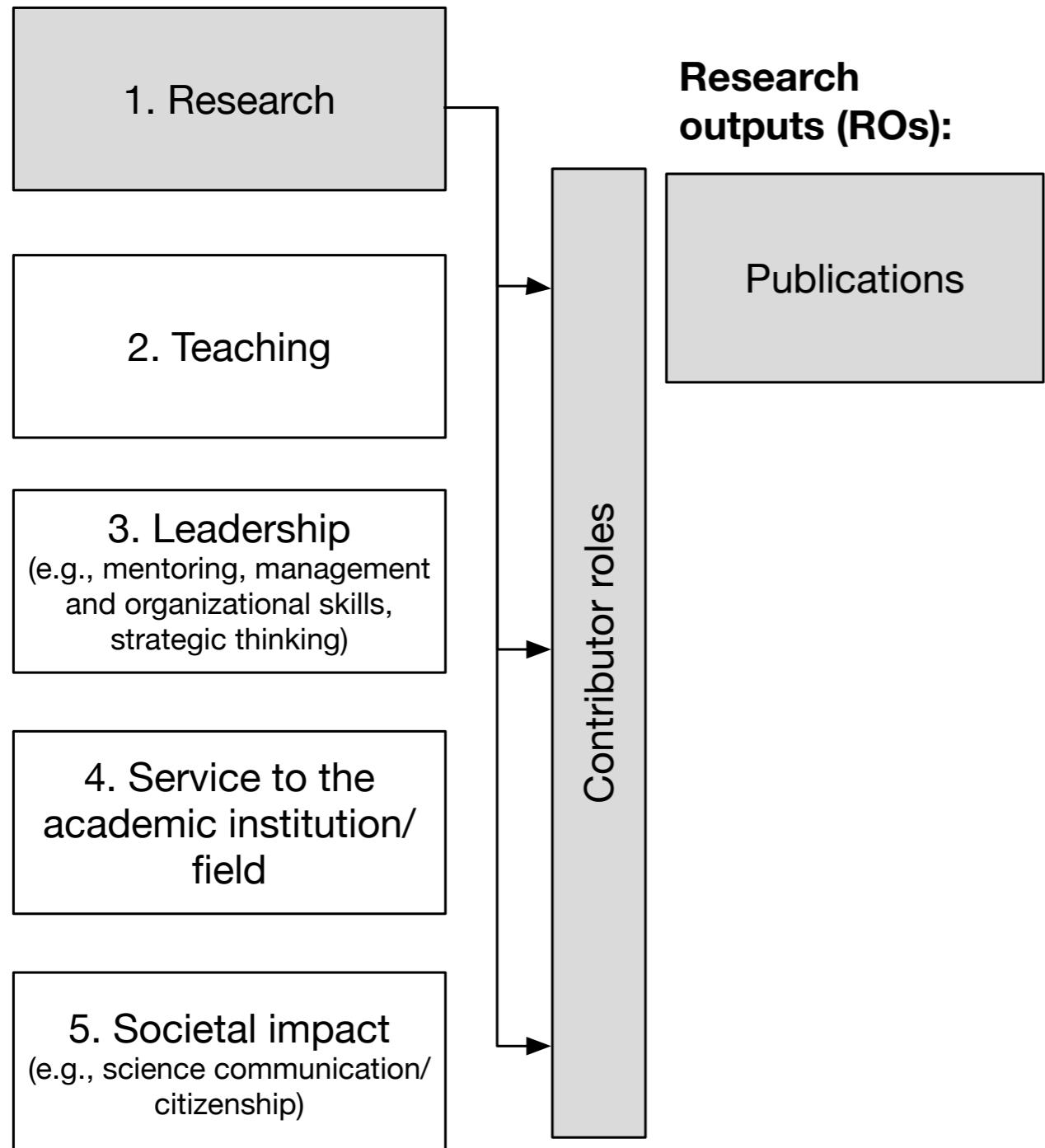
JIF: 24!

Citations:
2757!

My contribution:
Nearly zero.

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

Types of academic contributions:

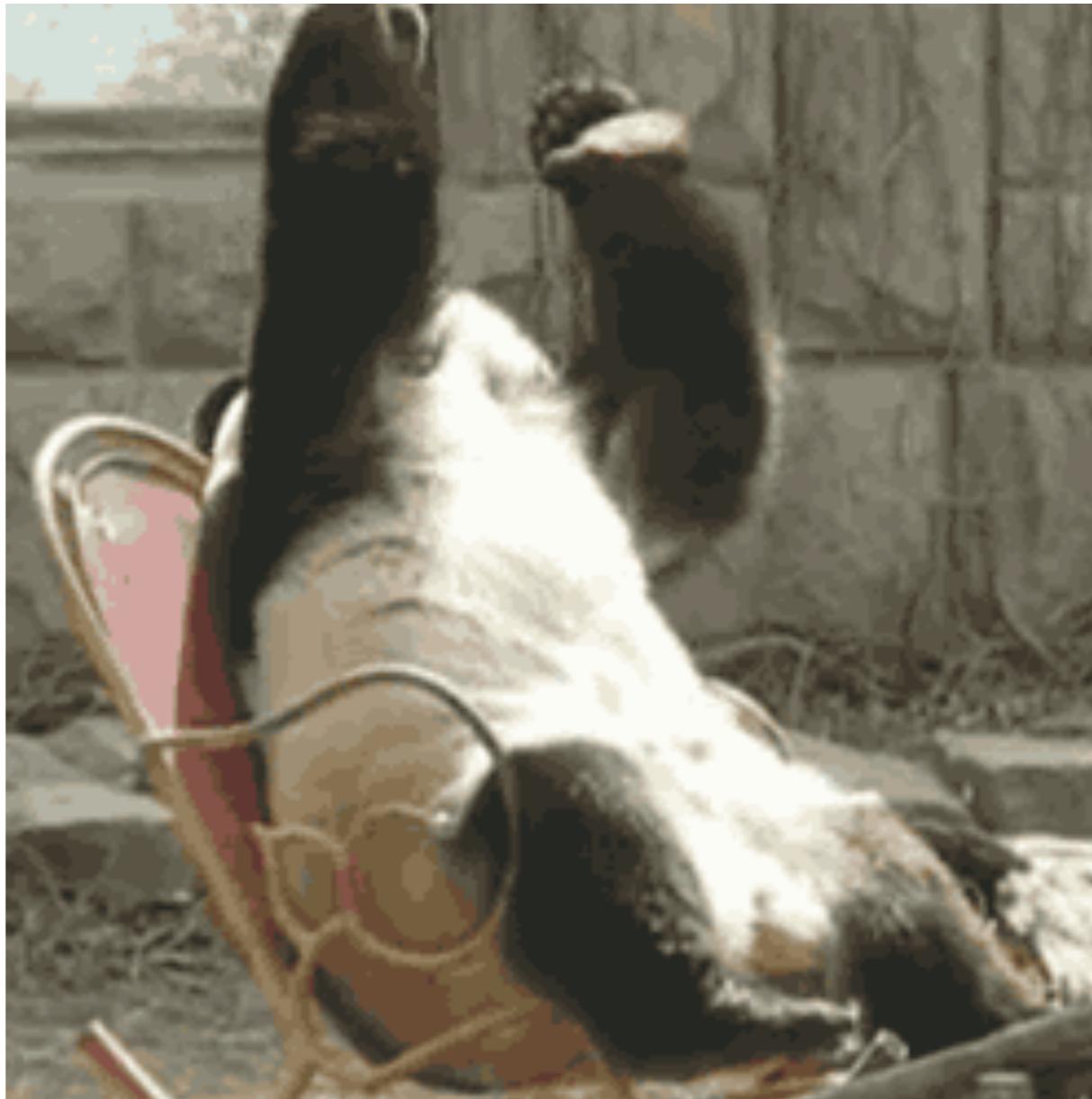


Quality over quantity



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

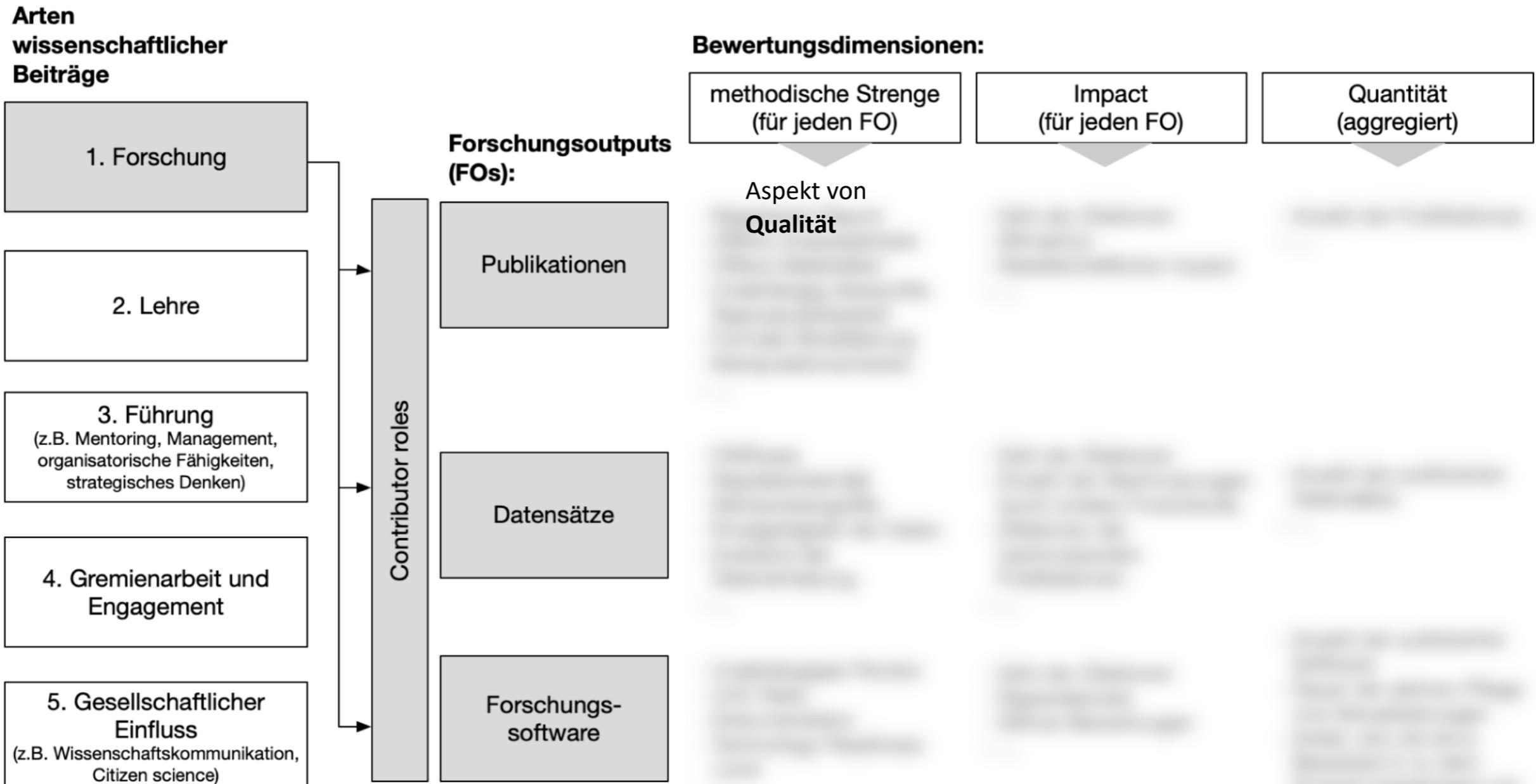
In this work, we use the number of citations as a proxy for quality,



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.

Research Quality Evaluation Scheme (RESQUE)



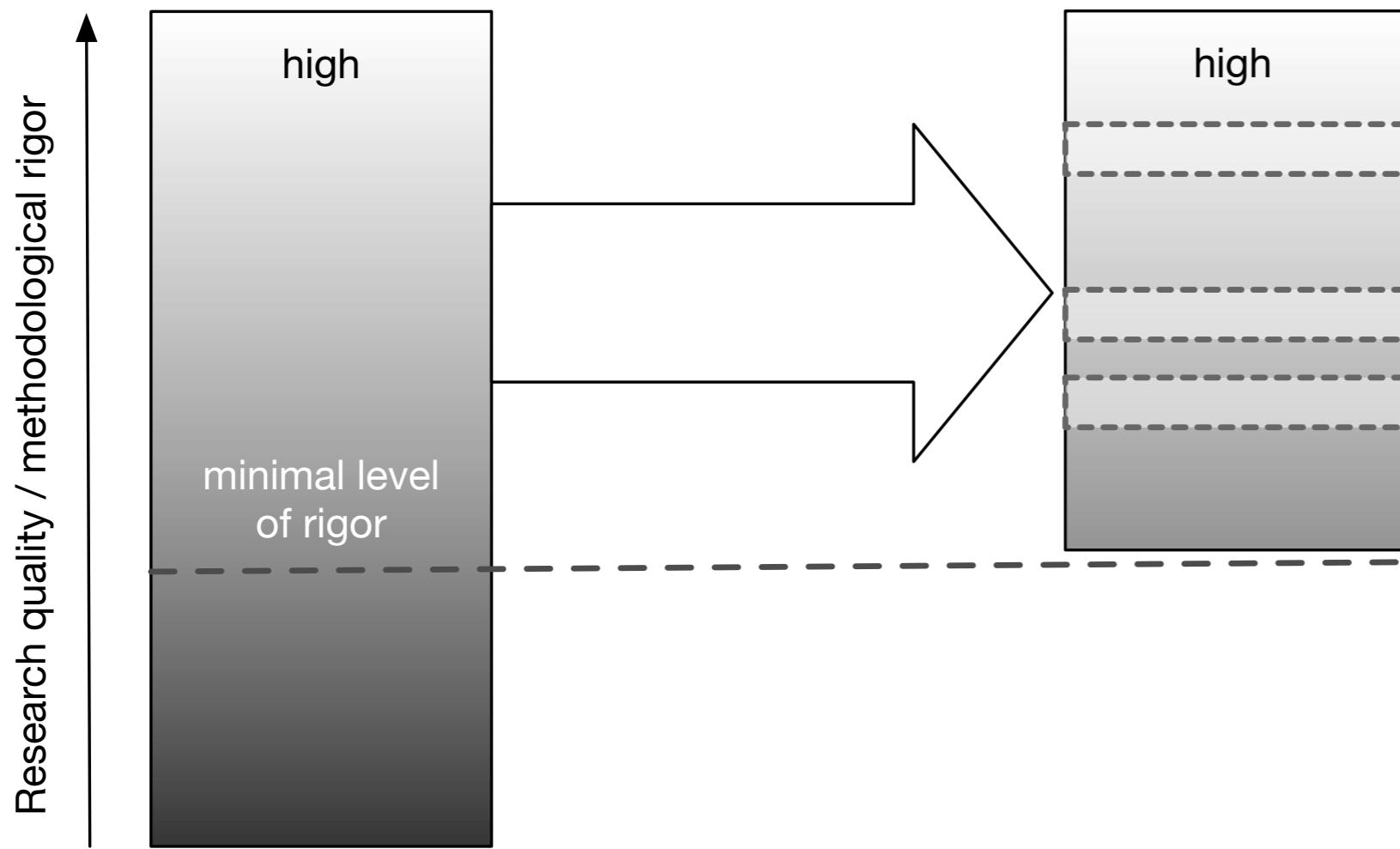
siehe Gärtner, A., Leising, D., & Schönbrodt, F. D. (2023). Empfehlungen zur Bewertung wissenschaftlicher Leistungen bei Berufungsverfahren in der Psychologie. *Psychologische Rundschau*, 74(3), 166–174. <https://doi.org/10.1026/0033-3042/a000630>; sowie <https://doi.org/10.31234/osf.io/rgh5b> und <https://doi.org/10.31234/osf.io/5yexm>

5. A two-phase hiring process

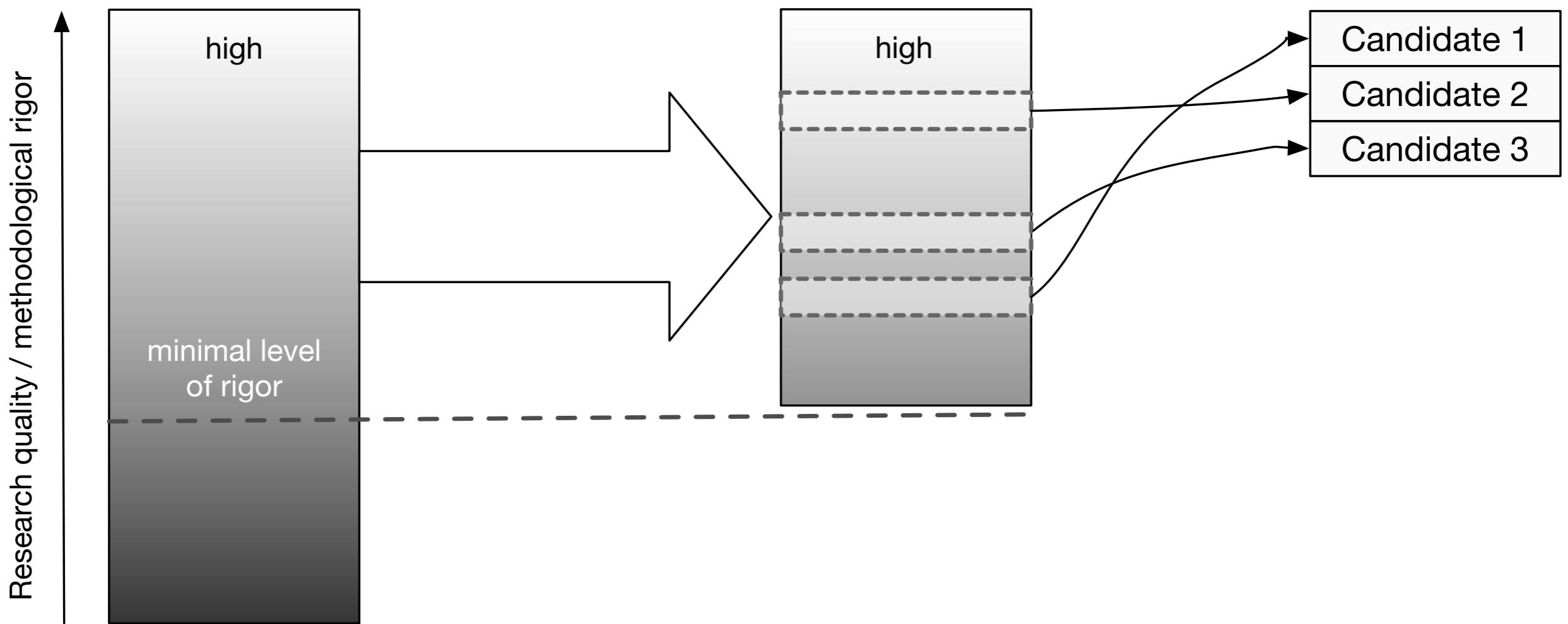
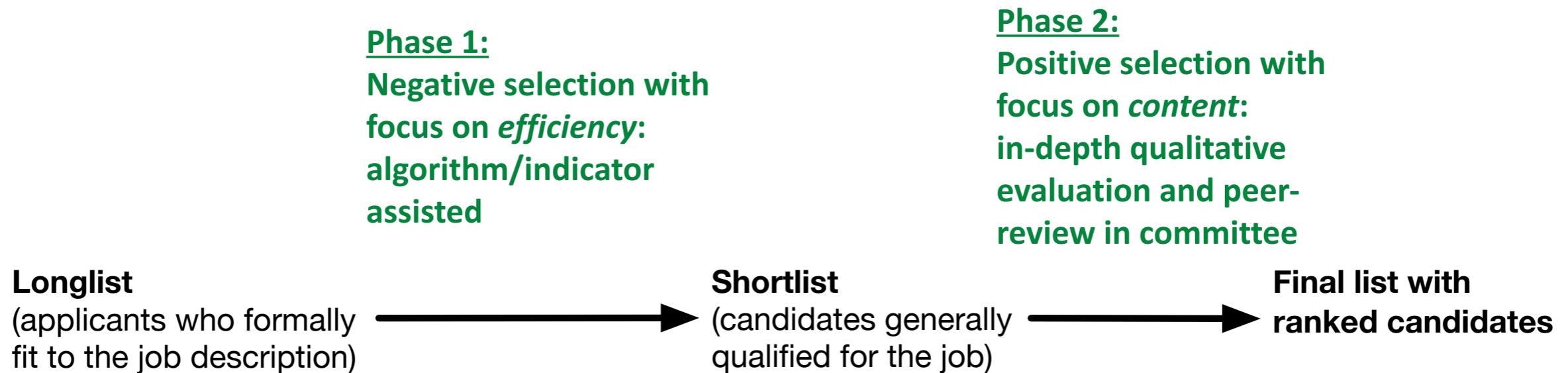
Phase 1:
Negative selection with
focus on *efficiency*:
algorithm/indicator
assisted

Longlist
(applicants who formally
fit to the job description)

Shortlist
(candidates generally
qualified for the job)



5. A two-phase hiring process



Scoring-Prinzipien in Phase 1

- Grundprinzipien:
 - Es gibt viele Arten, Wissenschaft zu machen. Aber was man macht, sollte man gut machen.
 - Bsp: Man muss nicht theoriegeleitet arbeiten - aber wenn man es tut, sollte es gut sein.
 - In Phase 1 messen wir „fast and frugal indicators“, die möglichst objektiv erfassbar sind.
 - Das Fokus liegt auf basalen *Hygienefaktoren*. (Wir versuchen erst gar nicht, Innovation etc. zu messen).
 - Man sollte nicht bestraft werden, wenn man Punkte prinzipiell nicht erhalten kann
- Relative Rigor Score
 - „POMP“: Percentage of maximum points
 - Wenn ein Indikator begründbar „not applicable“ ist, dann wird der Maximalscore entsprechend reduziert und man kann immer noch theoretisch 100% erreichen.

The RESearch QUality Evaluation Framework (RESQUE)

The RESQUE Fra...

 Get your RESQUE profile

 For hiring/tenure committees

 Publications

 Evaluation Projects

 Team / Contribute

 Technical documentation



RESQUE: The Research Quality Evaluation scheme for psychological research

The **Research Quality Evaluation (RESQUE)** framework provides recommendations for a responsible research assessment that does not rely on flawed metrics such as the journal impact factor or the h-index.

In alignment with the principles of [CoARA](#), this approach acknowledges diverse academic contributions, prioritizes the quality of work rather than its volume, and integrates qualitative peer assessment with the responsible use of quantitative indicators. Primarily designed to assist hiring and tenure committees, it emphasizes that the indicators and algorithmic methods serve as tools to support, not replace, decision-making processes. By automating the generation of relevant candidate information, this approach enhances the effectiveness of human expertise in evaluating potential hires and tenure candidates.

RESQUE provides objective quality and impact indicators for three types of research outputs:

- a. publications of empirical studies,
- b. published data sets and
- c. research software.

It is primarily developed for the field of psychology, but might be easily transferred to neighbouring empirical

News

See all news [here](#).

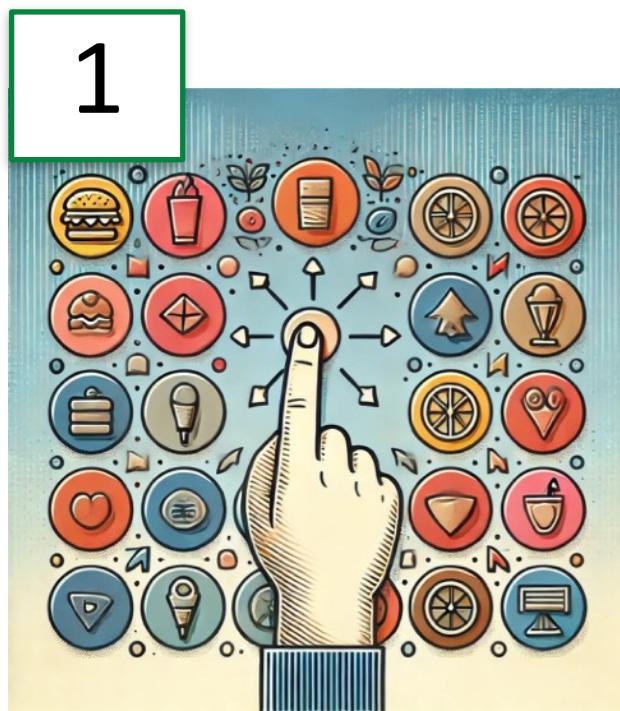
 2024-03-21:

[Einstein Foundation Award](#)

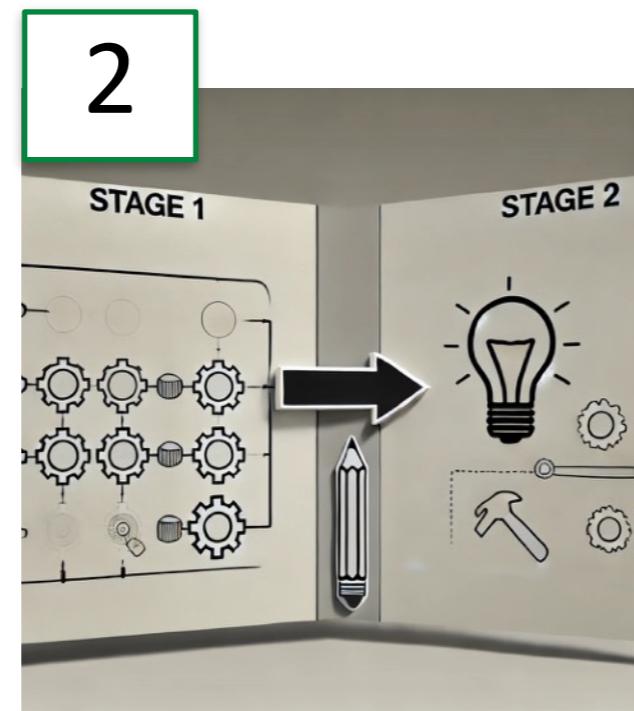
Anne Gärtner received the Einstein Foundation Early Career Award 2023 for the project on Responsible Research Assessment. The project is based on RESQUE and aims to further...

<https://resque-framework.github.io>

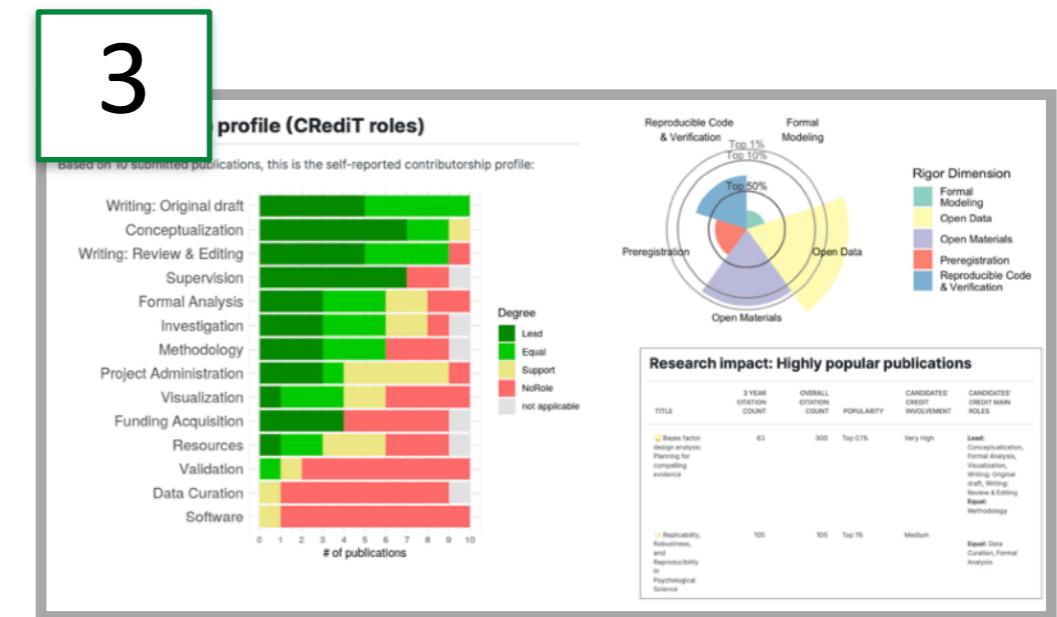
What is *RESQUE*? Three modular elements



1



2



3

✓ A set of evaluated indicators for aspects of research quality.

✓ A suggestion for a two-stage process and weights for indicators

👉 „Menu approach“: Escape h-index and JIF by choosing valid indicators that suit your research assessment scenario.

👉 Change the importance of indicators depending on your use case.

✓ A set of tools that can be used (optionally):

- Data entry for applicants
- Profile creation for CVs
- Interactive comparative overview for committees

👉 Open source, free to use, privacy aware.
Of course you can also include the selected indicators in your existing recruitment system.

Demo Collector-App:

<https://resque-framework.github.io/collector-app/>

[Note: The RESQUE app is in beta stage and might change in the near future. If you want to use it in practice, please [contact us](#).]

RESQUE [Clear](#) [Load](#) [Save to file ...](#)

10 of 10 slots used, you can't add more publications.

Author / Metadata
Schönbrodt

Publication 1
Machine Learning and Risk Assessment: Random Forest Does Not Outperform Logistic Regression in the Prediction of Sexual Recidivism

Publication 2
Big little lies: a compendium and simulation of *p*-hacking strategies

Publication 3
Replicability, Robustness, and Reproducibility in Psychological Science

Author / Metadata

Last name of applicant
Schönbrodt

First name
Felix

Year of PhD / Dr.
2010

Years that are subtracted from active academic life (e.g. due to child care). You can enter fractional years, e.g. 0.5.
0

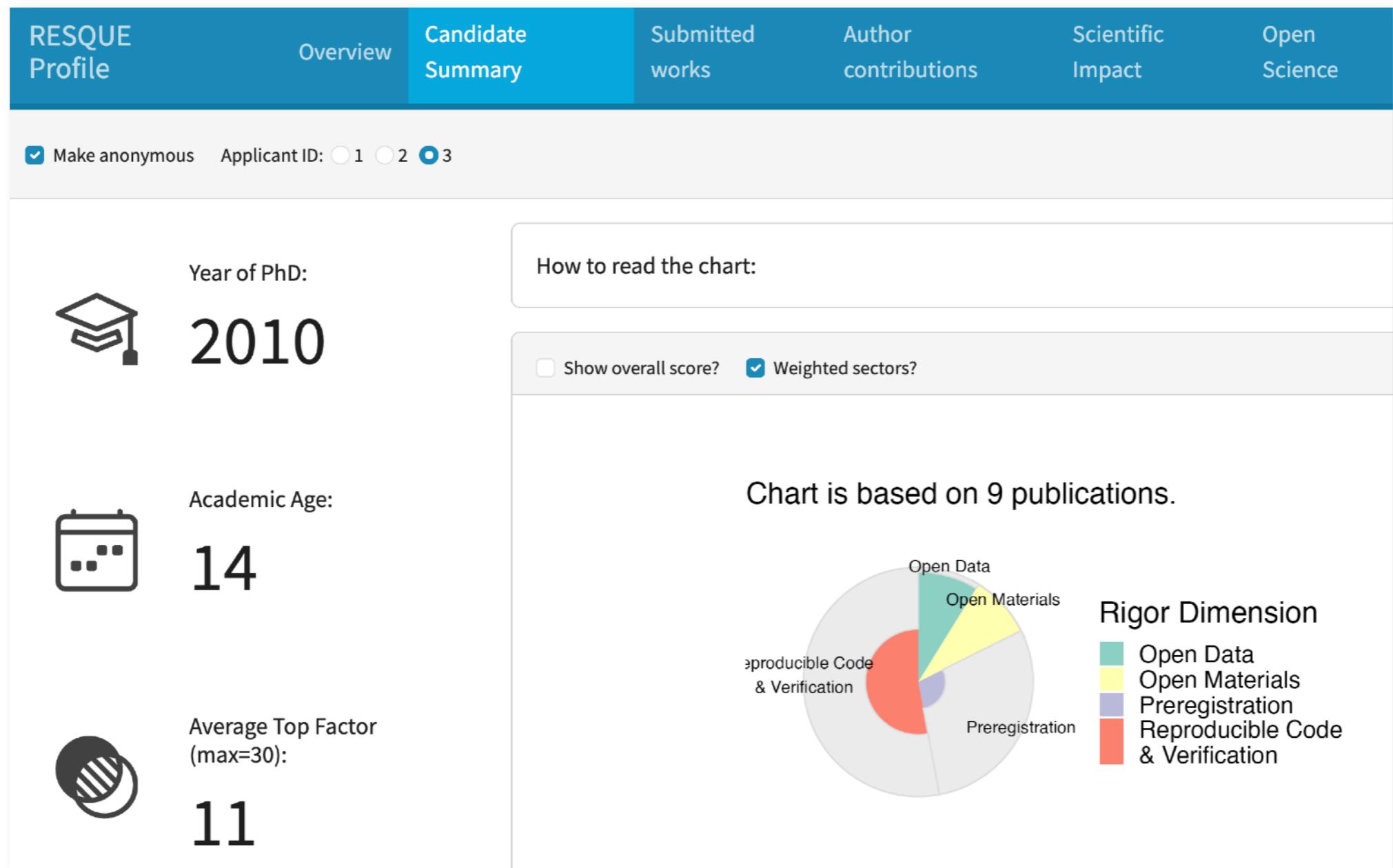
ORCID identifier (as full link, e.g. <https://orcid.org/0000-0002-1825-0097>)
<https://orcid.org/0000-0002-8282-3910>

Who enters the data?

I am the applicant & a co-author of these research outputs

Demo Profile-App:

R-package RESQUER: <https://github.com/RESQUE-Framework/RESQUER>



RESQUE: Umsetzung vor Ort

- Diskussion vor Ort initiieren:
 - Wie wird wissenschaftliche Leistung gemessen? In ...
 - Leistungsorientierter Mittelvergabe (LOM)
 - Tenure-track-Verfahren und Habilitationen
 - Berufungskommissionen
 - Kritische Reflektion über (möglicherweise) bisher genutzte Metriken (z.B. JIF, *h*-Index, Anzahl Publikationen)
 - Mögliches Ziel: Ein generelles [Mission Statement](#) zum Thema „Bewertung von Forschungsleistung“ im Institut
 - → **Wir bieten Impulsvorträge an, um diese Diskussion zu starten**
- Pilot-Einsatz von RESQUE in Berufungskommissionen
 - Die Elemente auswählen, die vor Ort passen; Gewichtung festlegen
 - Zeit sparen durch technische Umsetzung:
Eingabemaske für Bewerber:innen + übersichtliches und vergleichendes Profil
 - → **Wir unterstützen gerne bei der Umsetzung & Evaluation**