Use of Good Science criteria in hiring professors

Daniel Leising, Felix Schönbrodt, Anne Gärtner

The problem

- Metrics of research productivity play a key role in hiring for positions in academia, especially professorships
- They may have been introduced to provide some measure of fairness and objectivity
- BUT: The most commonly used metrics are simply too flawed to stay in use
  - Number of papers published (especially first-authorships)
  - Number of citations
  - h-Index
  - Journal Impact Factors
  - Grant Money acquired

The problem

- These metrics reward „getting published“, „getting cited“ and „getting funded“ in a system in which actual *quality controls* are simply too weak
- They reward *minimal effort* and often reflect *eminence* more than anything else
- Actually *gaming* these metrics is very easy and common

The problem

- The result of this is a research literature with an unacceptably low signal-to-noise ratio (a.k.a. „replicability crisis“)
- The actual goal of scientific activity is to provide robust incremental knowledge gain with the potential to contribute to the welfare of society – which *funds* said activity
- Given the current use of metrics, this goal plays too little of a role, as compared to

  (a)The career-interests of individual researchers, and

  (b)The self-preservation and growth interests of „research“ institutions such as universities

The possible solution

- As idealists who really, actually, seriously continue to believe in the intrinsic promise and worth of science, what shall we do?

- Key insights:
  - Not everyone is equally well-qualified to work in science
  - Many people would like to work in science
  - Some kind of easy-to-use metric of relative research productivity will thus be needed

The possible solution

- Instead of rewarding the mainly quantative and far too proximal output of the research pipeline, directly reward features of research that make robust, incremental knowledge gain more likely!
- This should be done at the level of individual contributions (papers) and may then be aggregated for any given journal, faculty, or university…
- ….or researcher (in **hiring** procedures)

Current state of the project

A few months ago, several different groups started to push for the development of a concrete proposal as to how hiring processes may be improved:

- The leadership of the DPPD section of the German Psychological Society
- The leadership of the German Psychological Society, with its Open Science Committee and ist newly appointed committee on „Incentive Structure, Power Abuse and Scientific Misconduct“

- Insight: Let‘s work on this together!

Current state of the project

- Currently in development: Two-part proposal with

  (a) General principles

  (b) An actionable implementation example

- The latter contains many more specifications, which are often somewhat arbitrary and depend on people‘s values and priorities. But they are necessary, and making them **explicit** is certainly a step forward as compared to the current situation.

- First drafts by Anne Gärtner, Felix Schönbrodt and myself

- Next (soon): Publication / call for comments

Context

- The use of metrics is a crucial problem in academic hiring processes, but it is not the only one.
- Other relevant issues
  - The **lack of incentives** for good quality in hiring committee work
  - The **lack of resources** (especially time and personnel) for good quality hiring committee work
  - The **lack of basic diagnostic qualifications** among most members of hiring committees
  - The **lack of valid assessment tools** for personal integrity and leadership skills
  - The influence of **politics** (Who likes whom? Who is perceived as threatening competition? Who is part of which friendly or rival networks? Who will be helpful in increasing my own status within the faculty?)
- All of these will be ignored here, but may addressed by the DGPs amwf committee

A specific, actionable implementation

Two phases

First Stage

We simply check whether a candidate conducts research that even only has the **potential** to contribute to robust incremental knowledge gain

Second Stage

We deal with what that research is actually about, whether it is innovative, creative, useful etc.

A specific, actionable implementation

1. Applicants for professorships name up to ten publications that they contributed to in the course of the last ten years
2. Journal names are omitted
3. Open question: how to deal with relative contribution size

First Stage

A specific, actionable implementation

4. Applicants provide the following information for each of these ten papers
   - Is the data openly available? Where?
   - Is the data openly available in FAIR format?
   - Are reproducible analysis scripts openly available? Where?
   - Are all confirmatory analyses marked as such and were they pre-registered?
   - Are the analyses tied to a formal model?
   - Does the paper contain at least one replication attempt?
   - Was statistical power for detecting these effects at least 80%?
   - Was sample sized determined a priori?
   - Can study claim sample representativeness for persons, stimuli?
   - What is the paper‘s relative citation ratio?

First Stage

A specific, actionable implementation

5. Applicants are awarded specific point values for meeting any of these Good Science criteria
6. Points are summed for each applicant and used to decide who gets invited for an interview, either based on a minimum threshold, or just based on relative overall merit
7. Veridicality checks may be done in a spot-check manner at this stage, but will have to be done for all applicants who go on to the next stage
8. Possible Exception: Applicants working mainly in theory development may fill up to three slots in the table with theory papers. These will then not be assessed in the first stage but will play an important role in the second.

First Stage

Table 1. Evaluation scheme for publications.

Criterion	Paper 1		Points Paper 1
Paper (without journal name)	Leising, D., Scharloth, J., Lohse, O., & Wood, D. (2014). What types of terms do people use when describing an individual's personality?		
doi	10.1177/0956797614541285		
Author roles - who did what? (cf. CRediT roles)	Leising conceptualized the paper and performed all statistical analyses, Leising and Wood wrote it, Scharloth performed the psycholexical analyses, Lohse collected some of the data.		
Own data or data reuse (choose all that apply)	Original data collection	☑	
	Data reuse	☐	
Open data	Not available	☐	0.5
	Not applicable [provide explanation]	☐	
	Yes [provide doi or URL]	☑	
↳ **FAIR format**	Yes	☑	0.5
	No	☐	
Open reproducible scripts	Not available	☑	0
	Not applicable [provide explanation]	☐	
	Yes [provide doi or URL]	☐	
Pre-registration	Not available	☑	0
	Not applicable [provide explanation]	☐	
	Yes [provide doi or URL]	☐	
Pre-registration	Not available	☑	0
	Not applicable [provide explanation]	☐	
	Yes [provide doi or URL]	☐	
Formal modelling	Yes	☑	2
	No	☐	
Replication attempt	Not available	☐	0
	Not applicable [provide explanation]	☐	
	Yes [provide doi or URL]	☐	
	Registered Report	☐	
Sample size			
A priori Power analysis done	No	☐	0
	Not applicable	☐	
	Yes	☐	
Sufficient a priori statistical power (please evaluate)	Unclear because no a priori effect size was specified, BUT certainly sufficient for typical effect sizes in this field (around r = .20).		
Representative subjects	Yes	☐	0
	No	☐	
Representative stimuli	Yes	☐	0
	No	☐	
Relative Citation Ratio (RCR)			

A specific, actionable implementation

9. To keep the workload for hiring committees as low as possible, applicants themselves provide the respective information online. The point value computation is automatized.

10. To keep the workload for applicants as low as possible, we recommend establishing a central registry in which applicants only have to provide this info once (e.g., at ZPID). Any hiring committee may then simply tap into this info, pending the applicant‘s permission. Or it may even be made public.

First Stage

A specific, actionable implementation

11. Open data sets that have been used (at least once) by other researchers shall be listed and will earn an applicant points, too
12. Programming of publicly available software (e.g., R-)packages that may be used by other researchers shall also be listed and will earn an applicant points

First Stage

Template for Published Datasets

Table 2. Evaluation scheme for published datasets.

	Dataset 1		Points Data 1
Title			
Year of publication	-		
doi			
Author roles - who did what? (cf. CRediT roles)			
Data type(s)	questionnaire	☐	
	behavioral	☐	
	physiological/biological	☐	
	other:	☐	
Study mode	online	☐	
	laboratory	☐	0
	both	☐	
↳ FAIR format	Yes	☐	
	No	☐	0
Sample size			
Number of items / variables			
Citations / Reuse indicator (evaluate relative to the age of data set!)			
Merit / impact statement (narrative, max 150 words)			
	Sum Dataset 1:		**0**
	Sum Dataset Bewerby 1:		**0**

Template for Research Software incl. Reuse

Table 3. Simple evaluation scheme for research software.

	Research Software 1		URL	Comment
Title	R package RSA		https://CRAN.R-project.org/package=RSA	
Citation	Schönbrodt, F. D. & Humberg, S. (2021). RSA: An R package for response surface analysis (version 0.10.4). Retrieved from https://cran.r-project.org/package=RSA			
Short description	An R package for Response Surface Analysis			
Date of first full release	2013			Necessary to compute citations relative to age of software
Date of most recent major release	2020			Indicates whether software is actively maintained
Contributor roles	Design	☐		What has the applicant contributed? Check all that apply. Possible values: • Design • Debugging • Maintenance • Coding • Architecture • Documentation • Testing • Support • Management
	Debugging	☐		
	Maintenance	☐		
	Coding	☐		
	Architecture	☐		
	Documentation	☐		
	Testing	☐		
	Support	☐		
License	GPLv3			Is the software open source?
Scientific impact indicators				
Downloads or users per month	710 downloads / month		https://cranlogs.r-pkg.org/badges/RSA	
Citations	110		https://scholar.google.de/citations?view_op=view_citation&hl=de&user=KMy_6VIAAAAJ&citation_for_view=KMy_6VIAAAAJ:mB3voiENLucC	Evaluate relative to the age of software
Reusability Indicator				
Merit / impact statement (narrative, max 150 words)				
	Sum Software 1:	**0**		
	Sum Software Bewerby 1:	**0**		

A specific, actionable implementation

- When applying, applicants are also asked to provide a narrative explaining what their research is about, how it relates to other work in the same field, and how their own work (**limited to the ten papers nominated in the first stage!**) has already advanced scientific knowledge or at least has the potential to do so.
- This narrative information on actual research *content* only comes into play at the second stage of the selection process. It is used as a basis for in-depth discussions with the applicants about their work. One or two committee members will have to actually read a number (2-5) of an applicant‘s contributions and thus be particularly well-prepared for these in-depth discussions.

Second Stage

Two side-notes

The current idea is that a scheme like the one presented here will have to be used for the next few (5? 10? 15?) years, in order to accelerate the use of Good Science practices. Ultimately, however, we hope this will become unnecessary because the use of Good Sciences practice will have become the rule rather than the exception after some time.

Two side-notes

But….what about funding? Funding does not play a role in this proposal because the contributions that a person makes to scientific knowledge do not become any stronger with how much money had to be spent in order to make them. What‘s more, weak scientific contributions may never be compensated for by spending as much grant money as possible to make them.

The primary role that funding plays in many hiring decisions is owed to the need of the hiring institution for **overhead money** (i.e., a percentage of the grant given directly to the institution) in order to ensure its daily functioning and, possibly, growth. Given the systematic underfunding of German universities at present, we consider this a legitimate goal, but one that is largely unrelated to purely scientific goals. Grant money acquisition should thus not be disguised as a measure of scientifc merit any longer but openly acknowledged as an additional hiring criterion that is motivated by the institution‘s need for self-preservation.

Open Questions

- Shall „competently handles multiple testing issue“ be a criterion?
- How shall we deal with „academic age“?
- How to handle the relative size of a candidate‘s contribution to a paper (1st, 2nd and last authorships only)?
- Narrative statements for each paper or for all of them together?
- Compensate theory-specialists by dividing score for empirical papers by number and then multiply with ten?
- Use minimum threshold (e.g., at least 30% of maximum attainable points) or relative ranking to select short list candidates from the long list? The latter is certainly more realistic.
- **Please help us optimize this thing by giving us feedback**