



## challenges in the peer-review process

Peer review is at the heart of the processes of all of science. It is the method by which grants are allocated, papers published, academics promoted, and Nobel prizes won. Yet it is hard to define. And **its defects are easier to identify than its attributes**. Famously, it is compared with democracy: a system full of problems but the least worst we have.





## does peer review `work' at all?

systematic review of all the available evidence on peer review concluded that **`the practice of peer review is based on faith in its effects, rather than on facts'**.





# THE DEFECTS OF PEER REVIEW

## Slow and expensive

many journals, even in the age of the internet, take more than a year to review and publish a paper (peer-review process, resubmission, etc.)

money becoming an issue (even if many editors, journals or funding agencies do not pay for writing review – time consuming process instead of doing own's research) – open access policy?

## Inconsistent

Reviewer A: *'I found this paper an extremely muddled paper with a large number of deficits'*

Reviewer B: *'It is written in a clear style and would be understood by any reader'.*





# Biased

**Matilda effect:** in which papers from male first authors are evaluated to have greater scientific merit than papers from female first authors, particularly in male-dominated fields

**Matthew effect:** already-famous researchers receive the lion's share of recognition for new work (cf. 'To those who have, shall be given; to those who have not shall be taken away even the little that they have').

**„Oxbridge” effect:** the fame or quality of the authors' institution(s) affects the peer review process



**Interdisciplinarity** (especially in grant proposal assessment)

while different organizations often promote the notions of interdisciplinary research and "synergy effect", in practice, it's not so clear how to approach such research projects





**difficulties in deciding what to work on:** the mismatch between the open problems and skills/qualifications of the team

good research problems can be found only in the interval between trivial problems and intractable ones - interdisciplinarity adds to the difficulty here, as each of the different disciplines has its own threshold for triviality and criteria for „interestingness”

a "good" interdisciplinary problem - the one which requires a solution incorporating complex methodological aspects from more than one area?





**difficulty** attaining results because of communication problems between team members with different backgrounds

**difficulty** attaining results because of lack of team members with deep technical knowledge in more than one of the big areas

**difficulty** publishing results because of distinct standards of evaluation and academic rigor used in distinct research areas

**difficulty** publishing results because of mismatch between academic publication venues (which are discipline- or subdiscipline-specific), and the scientific results of the project (which cover multiple areas)





peer review is the gold standard in most academic research assessment exercises, but research has demonstrated that this method is not only ill-suited to evaluating interdisciplinary work and it undermines truly innovative approaches





interdisciplinary work is too often the product of “amateurism and intellectual voyeurism” and results in knowledge of “dubious quality”

The real problem - the standard measures of scientific productivity and quality represent the “disciplinary assessment of an interdisciplinary work”





Lamont et al. 2006, Mallard et al. 2009, Lamont 2009: 81 expert panelists from a dozen multidisciplinary fellowship competitions in the social sciences and humanities asked how they assess the quality of interdisciplinary research proposals and how collectively the panel negotiates fairness in selecting winners: “discipline-specific ways of producing theory and methods are still the bedrock of peer evaluation”





the key to procedural fairness in assessing the quality of interdisciplinary work is not in giving up disciplinary autonomy, but in knowing when to give one set of disciplinary standards priority over others in the context of assigning greater value to proposals with intellectual breadth

if you have good expert panelists, the good interdisciplinary proposals will be given the chance and send to the second stage of the evaluation





**Language (especially in SSH)** – mostly in grant proposal evaluations - while assessing previous research achievements, the English language publications have more chances to become better evaluated than those in other languages, even if the quality of the journal (or the publications itself) would imply the opposite (the use of bibliometrics would be particularly biased since many of the tools – databases - are monolingual)





although modern science in many cases takes place in a collaborative environment that brings together researchers from all over the world, most scientific research is published in English – such a dominance of English in scientific publishing is fairly recent: before the First World War, the scientific literature was a mix of German, French, and English - English has continued to dominate as the language of science after WWII, **but it doesn't apply to all disciplines!**





Increase of the internationalization of the peer-review and various forms of evaluations for academic purposes (research grants applications, scientific outcomes of the research performing institutions, achievements of individuals) poses a great linguistic challenge to the reviewers





because the level of language expertise varies among researchers, the manuscripts that are submitted to journals are sometimes written such that the language interferes with the reader's understanding of the material.

Peer reviewers are faced with several poor options when they receive a manuscript with language issues.

- Reject the paper until it has been edited by a native speaker.
- Attempt to provide a review based solely on the content of the paper, even though that content may not be clear.
- Edit part or all of the paper while also providing feedback on the content of the paper.





“The PI is without a doubt a major specialist on the subject matter of this proposal. But his English is rather helpless (a notable exception being the paragraph in the middle of p. 6, which is suddenly flawless, probably being borrowed from a carefully copy-edited publication by the author), which does not augur well for the chances of making the results of this research available via major international publications”





“The PI is a trustworthy scholar, but he has not found his way to the international stage yet”

“The applicant has not published in major journals as he publishes mostly in German”

“Profile of the applicant not developed at the international level, all publications in the Russian language. The applicant seems to have published in Belarus journals so far. I cannot assess the quality of the journals nor the topics of the published articles”.

“Please note that due to the nature of the PI’s research focus and publishing history—I was unable to fully evaluate his competence as a scientist. I am unfamiliar with philology and cannot read German”





Topics – risky topics/niche topics – both in assessment of publications and grant proposals

editorial peer review process has been strongly biased against 'negative studies', i.e. studies that find an intervention does not work and the research grant application review process has been biased against very risky/niche topics

difficult to concentrate not on the results of a study but on the question it was asking - if the question is important and the answer valid, then it must not matter whether the answer is positive or negative – but what when we can already predict the result?





## TRUST IN SCIENCE AND PEER REVIEW

abuse of peer review - research misconduct – use of privileged information

difficult question is whether peer review should continue to operate on trust?

some journals, including the *BMJ*, make it a condition of submission that the editors can ask for the raw data behind a study; the editors did so to discover that reviewing raw data is difficult, expensive, and time consuming





Peer review is easy to get manipulated: as an editor or a funder you may easily manipulate the process – the “right” choice of experts is the most obvious case

**Poland 2013:** “cooperative” of several prestigious professors: an informal group acting as experts and reviewing grant applications (including international specialists) – the funds should stream to the “right” institutions (misconduct/criminal offence)





peer review is a flawed process, full of easily identified defects with little evidence that it works. Nevertheless, it is likely to remain central to science and journals because there is no obvious alternative, and scientists and editors have a continuing belief in peer review.

**odd that science should be rooted in belief**





# Validity Issues of Peer Review

## Bias and Predictive Validity

Michael Ochsner

COST ACTION CA15137 WG1

# Outline

## ➤ Outline:

➤ Issues

➤ Bias in Peer Review

➤ Predictive Validity?

➤ Example of bias in peer review

# Validity Issues in Peer Review

- **Subjectivity**
- **Low interrater reliability**
- **Moderate predictive Validity**
- **Biases**

# Subjectivity

- Own opinion, dependent on state of mind
- Outcome and decision depends on the selection of the reviewers
- Outcome depends on the availability of specialists
- Consequence: low validity as there is no clear reasoning (“I know it when I see it”, Lamont, 2009); no learning, fairness issues (Thorngate, Dawes & Foddy, 2009); low intra-rater reliability (Ochsner, Hug & Daniel, 2017)

# Low interrater reliability

- Two reviewers come to a different conclusion for the same work
- Discussions about decision, horse-trading, tit-for-tat
- Issue: do we really want high interrater reliability?
  - High interrater reliability might indicate old-boy network, poor selection of reviewers, same school of thought
- See, e.g. Daniel, Mittag & Bornmann, 2007

# Moderate Predictive Validity

- Research that was judged not good enough was later found to be influential
- Identified by comparing citations to rejected and published papers or projects
  - Issue: being funded/published might already enhance influence → numbers conservative
  - Bornmann & Daniel, 2008

# Biases

- Conservative bias: Experts tend to prefer research of their own style
- Mainstream bias: higher chances of being lucky to fall on reviewers who understand
- Other attributes than quality important: Gender, approach, (world)region
- Daniel, Mittag & Bornmann, 2007; Mutz, Bornmann & Daniel, 2008

# Example: Bias in Peer Review on Impact

- Conference «Pathways to impact from SSH research» in Vienna 28/29. November 2018



- First prize with the maximum of points to the UK. Second prize to the UK.
  - Jokes on Brexit
  - Talks about importance of responsible research and responsible metrics by jury members

# Winner



# Winner



The food that was not here anymore,

# Research = White



Need/Impact = Black, Research = White



# Agenda of Impact is Populist

- **The project is certainly excellent and impactful, not racist**
- **The video does what it needs to do**
- **The jury did what it was supposed to do**
- **The public did what it was supposed to do: applaud**
  
- **What did we learn about the project in the video? Nothing.**
- **What did we learn about the real reasons for the Problem? Nothing.**
- **Apartheid? Colonialism? The pope forbidding condoms? Nothing.**
- **It is a propaganda video for UK research, research that provides solutions**

# Neo-Colonialism at its Best: Cultural Bias

- No single member of the jury identified the issue!
- No person I talked to was suspicious
- No single person publicly opposed
  
- Why?!
  
- The impact agenda pushes to impact stories like that
- Demonstrable impact shifts focus from research to take-up and “Unique Selling Point”
- Professionalisation of “proves” of impact → Dramaturgy

# Limits of Peer Review

- **This example shows that peer review comes with issues (see issues identified above):**
  - **Subjectivity**
    - Selection of reviewers influences result
  - **Interrater reliability**
    - High interrater reliability might also point to some bias: the reviewers were "all of the same feathers"
  - **Moderate predictive Validity**
    - Do the reviewers identify the projects with the highest potential of impact or do they select "the best story", best dramaturgy?
  - **Biases**
    - Reviewers act on behalf of the mission but might not go beyond
    - Reviewers might be blind to some issues not related to their specialities or cultural background

# Literature

- van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014). The selection of talent as a group process. A literature review on the social dynamics of decision making in grant panels. *Research Evaluation*, 23(4), 298–311. <http://doi.org/10.1093/reseval/rvu017>
- Bornmann, L. (2011). Scientific Peer Review. *Annual Review of Information Science and Technology*, 45(1), 199–245.
- Bornmann, L., & Daniel, H.-D. (2008). The Effectiveness of the Peer Review Process: Inter-Referee Agreement and Predictive Validity of Manuscript Refereeing at *Angewandte Chemie*. *Angewandte Chemie International Edition*, 47(38), 7173–7178. <http://doi.org/10.1002/anie.200800513>
- Bornmann, L., & Daniel, H.-D. (2010). Reliability of reviewers' ratings when using public peer review: a case study. *Learned Publishing*, 23(2), 124–131. <http://doi.org/10.1087/20100207>
- Bornmann, L., & Daniel, H.-D. (2010). The Usefulness of Peer Review for Selecting Manuscripts for Publication: A Utility Analysis Taking as an Example a High-Impact Journal. *PLoS One*, 5(6), e11344. <http://doi.org/10.1371/journal.pone.0011344>
- Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics*, 2(4), 280–287.
- Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS One*, 5(12), e14331. <http://doi.org/10.1371/journal.pone.0014331>

# Literature

- Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society Series a-Statistics in Society*, 174(4), 857–879.
- Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge: Harvard University Press.
- Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review. *Science, Technology & Human Values*, 34(5), 573–606. <http://doi.org/10.1177/0162243908329381>
- Mutz, R. D., Bornmann, L., & Daniel, H.-D. (2014). Testing for the fairness and predictive validity of research funding decisions: A multilevel multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian science fund. *Journal of the Association for Information Science and Technology*, 66(11), 2321–2339. <http://doi.org/10.1002/asi.23315>
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2015). Does Gender Matter in Grant Peer Review? *Zeitschrift Für Psychologie*, 220(2), 121–129. <http://doi.org/10.1027/2151-2604/a000103>
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2016). Funding decision-making systems: An empirical comparison of continuous and dichotomous approaches based on psychometric theory. *Research Evaluation*, 25(4), 416–426. <http://doi.org/10.1093/reseval/rvw002>
- Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5, 187–255.