Recommended Procedures for Effective Peer Review

Hal R. Arkes
Ohio State University
November 1, 2017

My Peer Review Combat Experience

- 4 years as National Science Foundation (NSF) program officer, co-presiding over 13 peer review panels, attended many more.
- Co-presided over peer review panels at EPA and Department of Education
- Member of National Institutes of Health's (NIH) Review of Grant Application Committee, which attempted to revamp the NIH peer-review system.
- Ditto for NSF.
- Field experiments on peer review performance
- Publications on peer review performance
- Etc., etc.

Early Data on Peer Review of Scientific Work Done by Scientists

- Cicchetti (1991): Summarized inter-rater agreement in 11 studies on peer review within the social sciences: +.29.
- Gottfredson (1978): correlation between reviewers' ratings and eventual number of citations: +.28
- Mahoney (1977) and other studies found that congruence of research findings with a reviewer's point of view SIGNIFICANTLY influenced the reviewer's rating. (Psychology)
- Lloyd (1990) found that gender of author and reviewer influenced ratings. (Psychology)
  - Male reviewers did not rate male- and female-authored manuscripts differently.
  - Female reviewers were much more likely to accept female-authored manuscripts (62%) than male-authored manuscripts (21%).
- 1994 GAO Report on peer review at NIH, NSF, NEH
  - Females more likely than males to be funded at NIH.
  - Males more likely than females to be funded at NEH & NSF
  - Huge problem of gold standard for quality of manuscript/proposal.
  - Females and males were appropriately represented on panels.

More Recent Data on Peer Review In Science and Medicine

- Broder (1993). No influence of either the gender of reviewer or the gender of principal investigator on evaluations of NSF economics proposals.
- Abrevaya & Hamermesh (2012). Same result.
- Wenneras & Wold (1997): In Swedish Medical Research Council ratings females were rated lower in competence than males controlling for many other factors (university affiliation, education, etc).
- Grant, Burden, & Breen (1997): W&W result not replicated in UK.
- I cannot find any studies on peer review of minorities' research.

Ross et al. (2006)

- 67,275 abstracts submitted to 2000-2004 American Heart Assn's annual scientific sessions.
- Last 3 years had "blinded" reviews (i.e., no author or institution information). First 2 years had open reviews.
- Blinded reviews resulted in fewer acceptances to U.S. institutions and to prestigious institutions.
- Gender of author was irrelevant.
- Usual gold standard problem.

Hemlin's (1999) Summary

- More consensus in "certain hard sciences [special physics]) and
- disagreement in other (general physics, medicine) and behavioral sciences"
- This difference may be a consequence of the MUCH higher acceptance rate of journal submissions in some topics than in others.
  - Astrophysics (91%) (not much "room" for bias to be manifested)
  - Sociology (13%) (plenty of "room" for bias)

Summary Thus Far

- I cannot find a study published in the last 22 years that shows any gender bias
  - when scientists do the evaluation
  - of scientific materials (proposals, manuscripts)
  - in an English-speaking environment

Government Accountability Office (GAO) Complained in 1994 About Peer Review in 3 Federal Agencies

- Use of unwritten or informal criteria
- Lack of calibration among reviewers: Some were lenient and some were strict. Cole et al. (1981) found that the variance in NSF proposal ratings attributable to the reviewer was twice the size of the variance attributable to the proposal!!
- Many embarrassing examples: For example, a manuscript introducing the double-helix model of DNA was rejected by a journal.

Research On Problem #1: How Do NSF Panelists Use the 4 Official Criteria?

Table 1. *Use of four criteria by members of a dissertation award panel at the National Science Foundation*

	Criterion								
	Methodology		Student's training		Theory		Utility		Multiple
Panelist	Beta	Significance	Beta	Significance	Beta	Significance	Beta	Significance	*R*
1	.806	<.001	.108	<.01	.026	n.s.	.111	n.s.	.974
2	.332	<.01	.103	n.s.	.130	n.s.	.201	n.s.	.537
3	.381	<.001	.280	<.01	.120	n.s.	.244	<.05	.895
4	.445	<.001	.215	<.001	.076	n.s.	.334	<.001	.968

Note. The table shows results of an analysis in which the ratings of each panelist on the individual criteria were regressed on that panelist's overall rating.

Problem #2: Panelists Can Differ in Their Strictness

Table 2. *Hypothetical z scores illustrating harsh, lenient, and modal panelists' standards in using each of the five National Science Foundation rating criteria*

Score	Hypothetical panelist		
	Grim Reaper (harsh)	Polly Anna (lenient)	Professor Center (modal)
1 (*excellent*)	+2.3	+1.4	+2.0
2 (*very good*)	+1.7	+0.1	+1.0
3 (*good*)	+0.5	−0.9	0.0
4 (*fair*)	−1.2	−1.6	−1.0
5 (*poor*)	−1.7	−2.3	−2.0

Note. Entries represent *z* scores calculated for each panelist's ratings given to 20 or more proposals.

Simple Solution: Using Z-Scores Can Make a Difference, and It Can Level the Playing Field

Table 3. *Mean rating and z scores for individual criteria for every member of a Decision, Risk, and Management Science panel at the National Science Foundation*

	Panelist								
Score	A ($M = 3.35$)	B ($M = 2.65$)	C ($M = 2.95$)	D ($M = 2.77$)	E ($M = 3.26$)	F ($M = 2.82$)	G ($M = 3.36$)	H ($M = 2.48$)	I ($M = 2.39$)
1 (*excellent*)	2.67	1.88	1.81	1.59	2.60	1.73	2.36	1.51	1.40
2 (*very good*)	1.53	0.74	0.88	0.69	1.45	0.78	1.36	0.49	0.39
3 (*good*)	0.40	−0.40	−0.05	−0.21	0.30	−0.17	0.36	−0.53	−0.62
4 (*fair*)	−0.74	−1.53	−0.97	−1.11	−0.85	−1.12	−0.64	−1.55	−1.63
5 (*poor*)	−1.88	−2.67	−1.90	−2.01	−2.00	−2.08	−1.64	−2.57	−2.64

Blackburn & Hakel (2006)

- Examined the ratings of 1,983 posters submitted to 3 professional conventions.
- Reviewers didn't use z-scores, but Blackburn & Hakel converted all of the ratings to z-scores.
- Between 17% and 20% of the referees' decisions would have been reversed had z-scores been used.

The BIG Issue: How Can You Increase the Validity of Evaluations?

- Definition of "Reliability":
  - Intra-rater: If you do multiple ratings of the same research, do you give it the same rating each time?
  - Inter-rater: Do multiple experts give the same rating to the same research?
- Definition of "Validity": Does an evaluation assess what it is supposed to assess?
- VERY important fact: Validity is constrained by reliability. So because validity is so difficult to assess, reliability is often used as a proxy.

Methodology at SMDM Convention

- 6 sessions at the convention.
- 33 oral presentations
- 83 raters
- Everyone had all of the criteria listed on the top of every evaluation sheet.
- I divided each audience into those who rated each criterion separately and those who gave only one overall holistic rating.
- For the former group I just averaged the composite ratings to get their “overall” rating.

Criteria at SMDM Convention

- 1. Significance. Is the topic significant? Does it concern a scientifically important subject or is it relevant for health policy?
- 2. Methods: Are the methods scientifically sound?
- 3. Results: Are actual results presented in enough detail and in an understandable way?
- 4. Conclusions: Do the conclusions follow from the results? Are they justified?
- 5. Innovation: Is there something innovative about the presented material?

Inter-Rater Reliabilities for Scientific Presentations at SMDM Convention

- Disaggregated Method = .44
- Holistic Method = .18 (truly pathetic)

- Reliability of ratings mathematically constrains the validity of ratings. So disaggregated ratings almost certainly foster more valid ratings than do holistic ratings, because disaggregated ratings have significantly higher reliability.

How Many Scale Points Should Be Used?

- NIH's former 150-point scale?
- NSF's 5-point scale?
- Cicchetti et al. (1985) showed that once a scale included more than 7 points, inter-rater reliability either dropped or did not increase. Therefore I suggest that no more than 7 points be used.

Keep Records To Expedite Future Evaluations

- Klahr (1985) examined the ratings given by a particular NSF panel.
- He found that proposals whose average rating was better than 1.5 were always funded.
- He found that proposals whose average rating was worse than 3.5 were never funded.
- Therefore there was no point in taking the panel's time to discuss them. Just make a default decision.

Anonymity Can Help

- To remove either bias or the appearance of bias, conceal the name and affiliation of the person submitting the manuscript/proposal.

Summary

- You can check on the actual use of the appropriate criteria if you have evaluators provide ratings on each of the criteria and provide an overall rating.
- Use z-scores to calibrate the evaluators. (Easy)
- Use disaggregated ratings. (Easy)
- Use no more than 7-point scales. (Easy)
- Consider anonymity.

Thank you.
Questions?