Reviewer expertise and engagement in grant peer review

2nd Canadian Open Science Conference – pre-conference webinar

Sam Harper and Arijit Nandi

2026-10-02



McGill | Department of Epidemiology, Biostatistics and Occupational Health

Grant peer review is suboptimal

CIHR Project Grant review process

Our interest is here →→→

1. Three assigned reviewers read and score

Assigned partly on self-declared expertise

2. The committee discusses → consensus score

Reviewers present; all other members are “panelists”

3. Every member then scores privately

Final scores must stay within ±0.5 of consensus

CIHR Project Grant Peer Review Manual, cihr-irsc.gc.ca/e/49564.html

The context

Scores decide who gets funded

Peer review moves billions of research dollars

Scores come from humans in conversation

Panel discussion can change what reviewers think

Engagement and expertise are differential

Some members read the application closely; most did not

Prior literature

Reliability is low

Inter-rater agreement is moderate-to-low

Equity concerns are longstanding

Evidence of disadvantage by gender, race/ethnicity, and career stage in how panels score applications

Discussion is rarely studied

"Strategic rating" and discussion dynamics can *worsen* inconsistencies and lead to bias

Thorngate, Dawes, and Foddy (2011); Tamblyn et al. (2018); Derrick and Samuel (2017); Witteman et al. (2019); Ginther et al. (2011); Guthrie, Ghiga, and Wooding (2018); Pier et al. (2017); Fogelholm et al. (2012)

Ch: Other comments? (pause) So with that, let's hear our new scores?

MP: So I'll move to a four.

Ch: Secondary?

CV: Uh, I'll move to a four also.

Ch: Dr. Joshi?

GJ: I had four to begin with and I'll stay there.

Ch: Anyone outside that ra—these are pretty serious concerns that were raised. Four is a very high score.

JR: Yeah.

CV: Yeah mine, actually go to a five.

AA: (laughter)

Ch: Okay.

GJ: I'll go to five.

MP: I'll go to five.

Ch: Let's go again.

Pier et al. (2017), Figure 6

Approach: Analyzing CIHR's funding data

Research questions

Aim 1: Expertise and engagement

Do reviewers vs. panelists, and high vs. low expertise, differ in how they deviate from consensus?

Aim 2: Whose scores move?

Do those patterns differ by applicant characteristics?

Aim 3: What if?

Exploring alternatives (weighting by engagement, partial randomization).

Elements of CIHR's funding analytics database

Indicator	Availability
Application	
Funding result	Extractable
Consensus score	Extractable
Resubmission status	Extractable (2023+)
Keywords / domain, # investigators, funding requested	In-house only
Applicant	
Prior funding success	Extractable
Gender, career stage	In-house only
Scientific productivity	To clarify
Member	
Role (reviewer vs. panelist)	Extractable
Self-declared expertise (High/Med/Low/Not enough)	Extractable
Final score, initial reviewer scores	Extractable
Prior funding success, conflicts of interest	Extractable
Gender	In-house only
Reviewer experience/tenure	To clarify

Source: email from CIHR's Funding Analytics Team, Dec. 4, 2025. "In-house only": not shared as extracts (mainly identifiability risk), but a CIHR analyst can prepare the data and run our code in-house. "To clarify": answered "unlikely" (scientific productivity) or "likely no, pending how we measure it" (reviewer experience).

Approach: bring the code to the data

Restricted application-level data

CIHR Funding Analytics holds it; we do not see identities

Code must be submitted remotely

Analysts run it and return results (no interactive debugging)

So we simulate first

Build, test, and pre-specify the whole pipeline on synthetic data

The simulated world

50 committees, **20-80** applications per committee

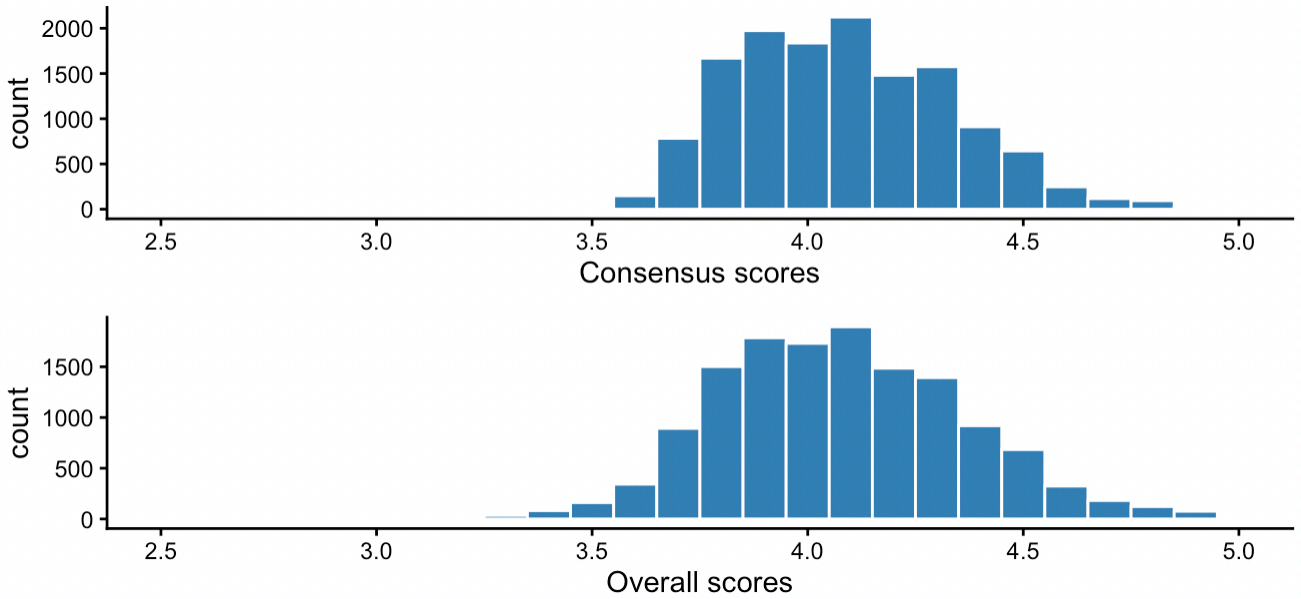
8-37 committee members

3 assigned reviewers, **15-16** applications discussed

Scores from truncated normal distribution

Modest treatment effects:

- **5pp** difference in deviation by engagement or expertise



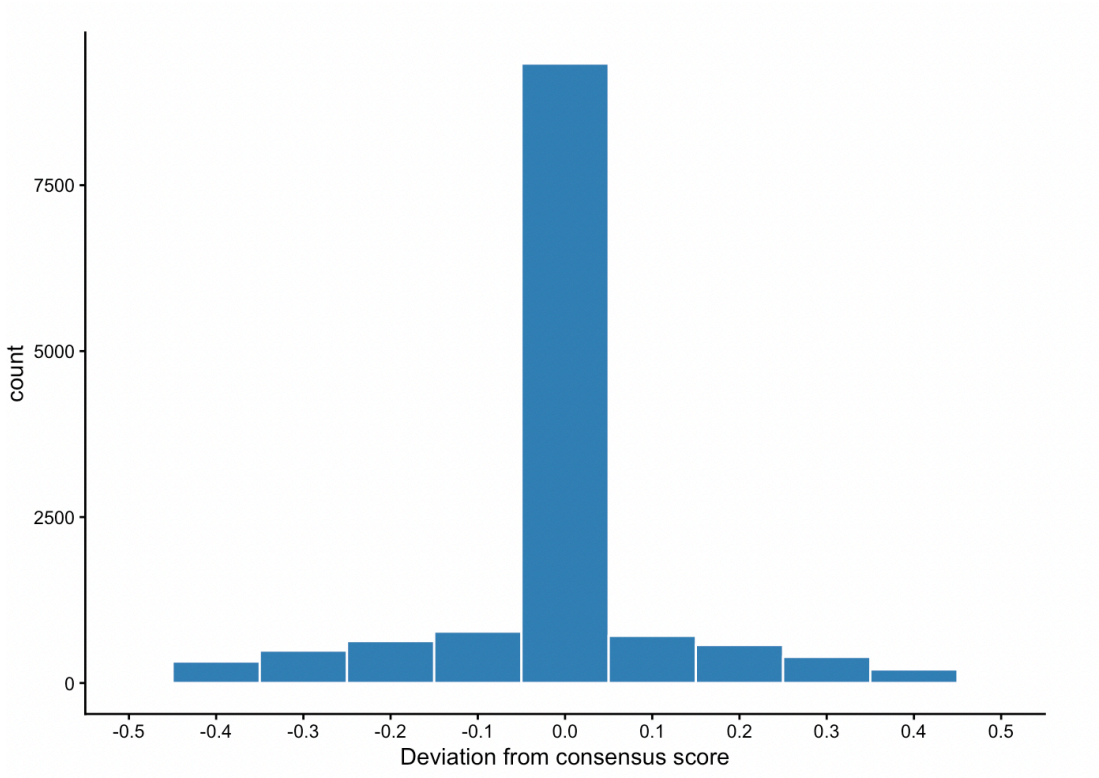
Outcome: d_{ijk} = final score – consensus score

Committee size/workload: CIHR Fall 2025 Project Grant committee rosters, cihr-irsc.gc.ca/e/54732.html; application pool range: CIHR's own stated 20-80 applications per committee

A two-part model for a “spike and slab” outcome

Outcome: d_{ijk} = final score – consensus score

Most likely adopt the consensus ($d = 0$)



Part 1: did they deviate at all?

Part 2: by how much, and in which direction, if they did?

$$E[d] = \underbrace{P(\text{deviate})}_{\text{Part 1: Bernoulli}} \times \underbrace{E[d \mid \text{deviate}]}_{\text{Part 2: ordinal}}$$

Bayesian multilevel models with committee, application and member effects

Both parts allow **role** and **expertise** to shift the *location* and, in Part 2, the *spread*

CIHR is actively reconsidering its own process

Strengthening peer review

Need to improve the quality of panel discussions and reduce burden

Process strain

Application volume up ~50% since 2023; Half of invited reviewers decline

Community frustrations

Concerns about reviewer workload and turnover, frustrations with virtual meetings

"Strengthening peer review together," CIHR, 2026-03-16, cihr-irsc.gc.ca/e/54669.html

Why dedicated metaresearch funding matters

Access

Partnership with the funder itself makes evaluating the funder's own process possible

Time to do it properly

Simulation, pre-specification and open code are slow and rarely rewarded in project grants

Open science with sensitive data

A reusable template for "open code, closed data" evaluation

Thank you

Sam Harper – McGill University

sbh4th@gmail.com

Funded by

Research on Research Joint Initiative: SSHRC, CIHR, Michael Smith Health Research BC

References

Derrick, Gemma, and Gabrielle Samuel. 2017. “The Future of Societal Impact Assessment Using Peer Review: Pre-Evaluation Training, Consensus Building and Inter-Reviewer Reliability.” *Palgrave Communications* 3 (1): 17040. https://doi.org/10.1057/palcomms.2017.40.

Fogelholm, Mikael, Saara Leppinen, Anssi Auvinen, Jani Raitanen, Anu Nuutinen, and Kalervo Väänänen. 2012. “Panel Discussion Does Not Improve Reliability of Peer Review for Medical Research Grant Proposals.” *Journal of Clinical Epidemiology* 65 (1): 47–52. https://doi.org/10.1016/j.jclinepi.2011.05.001.

Ginther, Donna K., Walter T. Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. 2011. “Race, Ethnicity, and NIH Research Awards.” *Science* 333 (6045): 1015–19. https://doi.org/10.1126/science.1196783.

Guthrie, Susan, Ioana Ghiga, and Steven Wooding. 2018. “What Do We Know about Grant Peer Review in the Health Sciences? An Updated Review of the Literature and Six Case Studies.” RAND Corporation.

Johnson, Valen E. 2008. “Statistical Analysis of the National Institutes of Health Peer Review System.” *Proceedings of the National Academy of Sciences* 105 (32): 11076–80. https://doi.org/10.1073/pnas.0804538105.

Pier, Elizabeth L., Joshua Raclaw, Anna Kaatz, Markus Brauer, Molly Carnes, Mitchell J. Nathan, and Cecilia E. Ford. 2017. “‘Your Comments Are Meaner Than Your Score’: Score Calibration Talk Influences Intra- and Inter-Panel Variability During Scientific Grant Peer Review.” *Research Evaluation* 26 (1): 1–14. https://doi.org/10.1093/reseval/rvw025.

Tamblyn, Robyn, Nadyne Girard, Christina J. Qian, and James Hanley. 2018. “Assessment of Potential Bias in