Local versus global inter-rater reliability for evaluating the internal validity of grant peer review: Considerations of measurement

Elena A. Erosheva[1], Patrícia Martinková[2], and Carole J. Lee[3]

[1]Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, Box 354320, University of Washington, Seattle, WA 98195.

[2]Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

[3]Department of Philosophy, University of Washington, Seattle, WA 98195

Introduction

Without a gold standard for what counts as a good outcome in a grant peer review process, inter-rater reliability (IRR) measures have been used as one approach to evaluating the internal validity of grant review processes. Evaluative disagreement among reviewers implies not only that decisions to fund proposals reflect the luck of the reviewer draw (Graves et al., 2011; Cole et al., 1981; Hodgson, 1997), but that the inherent credibility of peer review as an evaluative tool is up for question (Jayasinghe et al., 2001; Marsh et al., 2008). As such, Marsh et al. (2008) have described relatively low IRR rates in peer review to be the "most basic, broadly supported, and damning" evidence against peer review (p. 161).

There is a growing body of research on the IRR of grant peer review (e.g., Jayasinghe et al., 2001, 2003, Mutz et al 2012, Carpenter et al 2015), including a recent article that demonstrated "no agreement among reviewers regarding the quality of the applications" for NIH R01 grant proposals (Pier et al 2018). While the authors acknowledged that their study "included only high-quality grant applications" that were eventually funded and that it is not possible to "say whether these findings would generalize to an entire pool of applications," readers who are skeptical about the effectiveness of peer review may be tempted to infer that there is *no* reviewer agreement to be found across the entire spectrum of NIH R01 proposals.

In this work, we investigate the plausibility of estimates of zero inter-rater reliability for top-quality grant proposals. We show that, with less than 20 percent of top-quality proposals, zero IRR estimates are plausible with likelihood-based approaches even when the global IRR estimate is not zero. This is both due to range restriction and to difficulties in estimating between-group variance in the case of many small groups. Therefore, because of questionable validity, we recommend against estimating IRR for range-restricted samples in typical peer review settings. We conclude by discussing considerations of measurement when the interest is in making distinctions among top-quality proposals for funding decisions.

Illustration: IRR estimates for fractions of top-quality proposals

We use original peer review scores for all applications submitted to three rounds of one program to the American Institute of Biological Sciences (AIBS) to illustrate how IRR estimates change when varying fractions of top-quality proposals are considered. The American Institute of Biological Sciences (AIBS) provides independent, objective scientific peer review services. Our AIBS grant review data come from three rounds of applications to an ongoing intramural collaborative biomedical research program for 2014-2017. Most of the applications to this program are akin to NIH's R01 funding mechanism in that applicants can

request up to 3 years of funding with a maximum of $450,000 in direct costs. There were a total of 72 applications, each evaluated by three reviewers with areas of expertise closely matching those of the applications being evaluated. Individual reviewers provide scores for four application criteria (Innovation, Approach/Feasibility, Investigators and Significance) as well as the overall scientific merit score. Each of these criteria is scored on a scale from 1 (best) to 5 (worst); one decimal place is allowed in the scores. In this review mechanism, the final proposal score is the average of the three reviewers' scientific merit scores and is the primary factor in determining AIBS funding decisions. No panel discussion takes place in this review mechanism. For the purposes of our illustration, we focus only on estimating IRR measures for the overall scientific merit score. Research demonstrates that reviewers exhibit equal or worse IRR measures when scoring lower-level, proposal-related criteria such as originality, methodology, and scientific/theoretical merit (Jayasinghe et al., 2003).

We use original scientific merit review scores to illustrate changes in IRR estimates for various fractions of top-quality proposals. Thus, we order the proposals by the average rating of the assigned reviewers and estimate local IRR in succession for the top-quality 2, 3, 4, …, 71 proposals and global IRR for all 72 proposals. Results of our analyses are summarized in Figure 1. Panel A provides Maximum likelihood (ML) and Bayesian estimates for single-rater IRR and the associated uncertainty intervals for a given number of top-quality proposals, ordered by the average rating. As expected, when a fraction of top-quality proposals is considered, single-rater IRR estimates are low – including several cases of zeros – and there is considerable disagreement between ML and Bayesian approaches. However, considering all 72 applications, the ML estimate is 0.37 and the Bayesian estimate is 0.43, both being significantly different from zero. Panel B, analogous to Figures 5.4-5.5 presented by Raudenbush (2008) but with settings corresponding to the top 20% applications in AIBS data, provides the likelihood function for a key parameter is estimating the IRR – the between-group variance -- when the observed between-group variability is exactly equal to its expected value (black) and to a value which is just one standard deviation below (blue). The blue line therefore indicates a less favorable occurrence but one that could easily arise in practice. What is important is that the less favorable occurrence gives a ML estimate of zero for the within-group variance, thus also for the IRR.

In summary, we demonstrate that statistical inference for IRR estimates from small samples of similar quality proposals can mislead the unwary. As can be seen in Figure 1, one is quite likely to obtain small or even zero IRR estimates that poorly reflect plausible IRR values for the full range of applications.

Conclusion

Inter-rater reliability of grant review scores is one way to evaluate the internal quality of peer review. Our analysis demonstrates that it is fairly plausible to obtain estimates of exactly zero for IRR in a typical grant peer review setting when reviews for less than 20 percent of top-quality proposals are considered. This has several practical implications for studying the quality of peer review with IRR. First, we recommend against estimating IRR for range-restricted samples in typical grant peer review settings because of the questionable validity of those estimates. However, should this type of estimate be obtained, to avoid confusion, we suggest that, at a minimum, the word "local" must be included whenever one reports IRR estimates from restricted-range samples. This is because, as we illustrate in this paper using

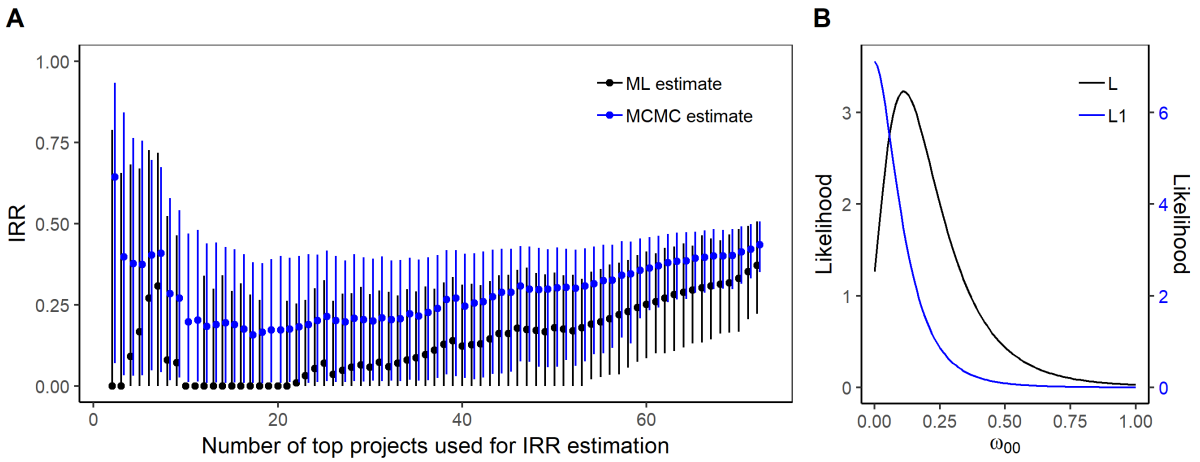

Figure 1, Panel A: Maximum-likelihood (black) and Bayesian (blue) estimates for single-rater IRR and the associated 95% bootstrap and MCMC uncertainty intervals calculated from a given number of top-quality proposals. Panel B: Likelihood function for between-group variance ω_{00} when observed between-group variability is exactly equal to its expected value (black) and when observed between-group variability is equal to a value that is one standard deviation below its expected value (blue).

grant review scores across the full range of proposal submissions, estimates of local inter-rater reliability from range-restricted samples are smaller than estimates of global inter-rater reliability, a point previously noted by Jayasinghe et al. (2001, p.350) and Jayasinghe et al. (2003, p.297) in the context of IRR estimation. This point is analogous to an observation made by Lindner and Nakamura (2015, p.5) about the important distinction between using range-restricted versus full-range data to measure the predictive validity of grant review.

In addition to technical estimation problems that are a danger to valid statistical inference for local IRR estimates, an important question for peer review researchers is whether range-constrained IRR is an appropriate construct for judging consistency in reviewer ratings. Is it valid to interpret range-restricted IRR when reviewers are asked to score grant proposals across the whole range of submissions?

Although global IRR is the only meaningful IRR-type characteristic of the corresponding peer review process when reviewers' task is to discriminate among all possible grant submissions, in the current funding climate when funding rates can be less than 10 percent of submissions, the question of whether peer review is able to differentiate among top quality proposals is important. As funding scarcity continues, some reject the idea that peer review is capable of distinguishing superlative from excellent proposals (Pier et al., 2018; Fang et al., 2016) – that, in its current form, the process of peer review is asked to work "at a level of discernment that exceeds the 'resolving power' of the evaluation instrument" (Chubin et al., 1990). However, as we argue here, one cannot answer definitively the question of the 'resolving power' of the ratings for top-quality applications by post-processing data from peer review settings where reviewers' evaluative task is to assess proposals across the full range of quality. A better procedural and substantive separation of the global (rating all possible applications) from the local (rating only top-quality applications) could be one way forward.

Acknowledgements: This work was supported by the NSF grant 1759825 awarded to E.A.E. (PI) and C.J.L. (co-PI). P.M. was supported by COST Action TD1306, the EU Framework Programme Horizon 2020. The authors would like to thank Stephen Gallo and the American Institute of Biological Sciences for providing grant review data.

References:

Cole, S., G. A. Simon, et al. (1981). Chance and consensus in peer review. *Science* 214(4523), 881–886.

Chubin, D. E., E. J. Hackett, and E. J. Hackett (1990). *Peerless science: Peer review and US science policy.* Suny Press.

Fang, F. C., A. Bowen, and A. Casadevall (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *Elife* 5, e13323.

Graves, N., A. G. Barnett, and P. Clarke (2011). Funding grant proposals for scientific research: Retrospective analysis of scores by members of grant review panel. *BMJ* 343, d4797.

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology* 50(11), 1189–1195.

Jayasinghe UW, Marsh HW, and Bond N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis 23*(4), 343–364.

Jayasinghe UW, Marsh HW, and Bond N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 166*(3), 279–300.

Lindner, M. D. and R. K. Nakamura (2015). Examining the predictive validity of NIH peer review scores. *PLoS ONE* 10(6), e0126938.

Marsh HW, Jayasinghe UW, and Bond NW. (2008). Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability. *American Psychologist. 63*(3), 160–68.

Mutz, R., L. Bornmann, and H.-D. Daniel (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS ONE 7*(10), e48509.

Pier E, Brauer M, Filut A, Kaatz A, Raclaw J, Nathan M, Ford C, Carnes M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proc Natl Acad Sci USA 115*(12), 2952–2957

Raudenbush SW. (2008) Many small groups. Handbook of Multilevel Analysis, eds de Leeuw J, Meijer E, Goldstein H (Springer, New York, NY), 207–236.