BMC Research Notes

COMMENTARY

Open Access

Reimagining peer review as an expert elicitation process

Alexandru Marcoci[1*], Ans Vercammen[2,6], Martin Bush[3], Daniel G. Hamilton[3], Anca Hanea[3,4], Victoria Hemming[5], Bonnie C. Wintle[3], Mark Burgman[6] and Fiona Fidler[3]

Abstract

Journal peer review regulates the flow of ideas through an academic discipline and thus has the power to shape what a research community knows, actively investigates, and recommends to policymakers and the wider public. We might assume that editors can identify the 'best' experts and rely on them for peer review. But decades of research on both expert decision-making and peer review suggests they cannot. In the absence of a clear criterion for demarcating reliable, insightful, and accurate expert assessors of research quality, the best safeguard against unwanted biases and uneven power distributions is to introduce greater transparency and structure into the process. This paper argues that peer review would therefore benefit from applying a series of evidence-based recommendations from the empirical literature on structured expert elicitation. We highlight individual and group characteristics that contribute to higher quality judgements, and elements of elicitation protocols that reduce bias, promote constructive discussion, and enable opinions to be objectively and transparently aggregated.

Keywords: Peer review, Expert elicitation, Wisdom of the crowd, Anonymity, DELPHI

Introduction

Trust in the good judgement of reviewers with relevant qualifications, experience, and scientific skill is at the heart of peer review. Editors value reviewers' expertise, and multiple studies have found a strong correlation between editorial final decisions and reviewers' judgements [1–3]. However, human judgement (even experts') is often flawed, susceptible to conscious and unconscious prejudices, misunderstandings, and gaps in knowledge. It should therefore be unsurprising that peer review can also be biased [4]. Peer reviewers have been shown to overlook methodological flaws and statistical errors [5–7], avoid reporting suspected instances of fraud [8] and commonly reach a level of agreement barely exceeding what would be expected by chance [9]. Recent studies have also exposed the extent of gender bias in peer review [10] and questionable editorial protocols that lack transparency [11]. Despite the wide range of issues, the debate over whether the system is irrevocably broken has not been settled. What is clear is that more work is needed to understand how journal peer review functions, to identify pressure points and improve its efficacy as a gatekeeping mechanism for high quality science [12].

Contemporary peer review is predominantly a journal-organised, pre-publication quality assurance activity wherein independent and (often) anonymous reviewers provide their opinion on the suitability of submissions for publication [11, 13, 14], with reviewers being prompted with open questions or with several criteria that they should consider when making judgements [14]. Over the last two decades, editorial procedures have begun gradually to diverge from these conventional features, although the advances are slow and subject to criticism about time delays and undue influence over the process [13, 15]. A small minority of journals are experimenting

*Correspondence: alexandru.marcoci@gmail.com
[1] Centre for Argument Technology, School of Science and Engineering (Computing), University of Dundee, Dundee, UK
Full list of author information is available at the end of the article

with innovative peer review models that encourage dialogue between invited reviewers (between 2 and 8% of journals) [11, 14, 16]. While these more collaborative approaches are a promising development, they have not solved critical issues around limited participation, reviewer coherence or accountability. We believe that additional structural changes are required. To provide actionable recommendations for a fairer, accountable and more transparent system, we start from the observation that peer review should be treated like a structured expert elicitation process, applying tested protocols that incorporate research from mathematics, psychology, and decision theory to mitigate biases, and enhance the transparency, accuracy, and defensibility of the resulting judgements. This can demonstrably improve the quality of expert judgements, especially in the context of critical decisions [17–21]. In what follows we outline the hypothetical benefits of applying structured expert elicitation principles in journal peer review. In the Outlook section, we reflect on the challenges ahead.

Main text

Peer review as a structured elicitation process

Our recommendations are based on our collective experience developing and implementing the IDEA protocol (Investigate—Discuss—Estimate—Aggregate, Fig. 1) for structured expert elicitation in diverse settings including conservation, intelligence analysis, biosecurity, and, most recently, for the collaborative evaluation of research replicability and credibility [22–28]. This Delphi-style protocol has been shown to facilitate accurate predictions about which research findings will replicate [29] by prompting experts to investigate and discuss the transparency and robustness of the findings in a structured manner.

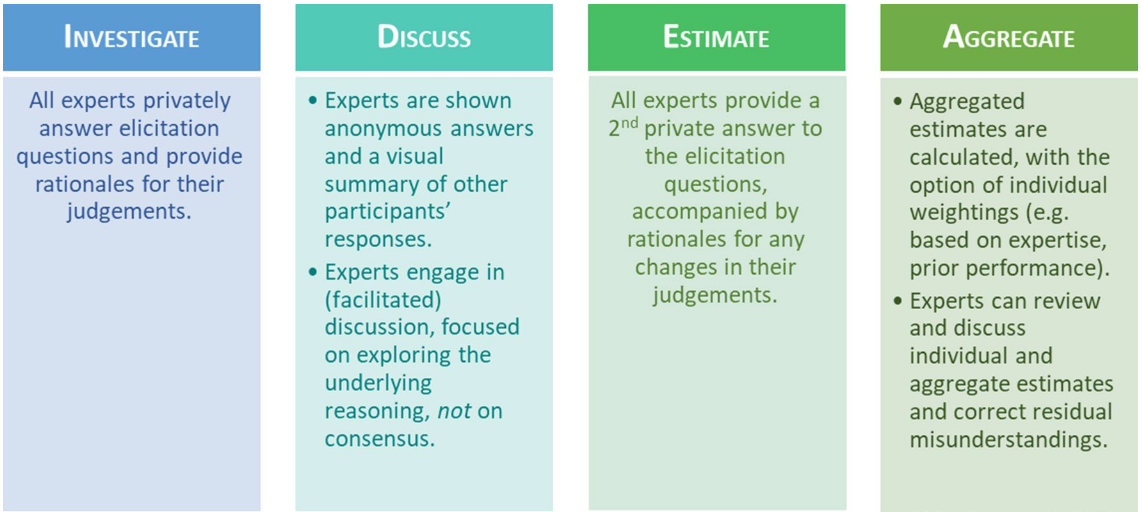
In the following sections, we outline five recommendations focusing on individual and group characteristics that contribute to higher quality judgements, and on ways of structuring elicitation protocols that promote constructive discussion to enable editorial decisions that represent a transparent aggregation of diverse opinions (Fig. 2).

Elicit diverse opinions

One of the crucial phenomena that underpins most expert elicitation protocols is the wisdom of the crowd effect [30], a statistical phenomenon where random errors associated with independent judgements cancel each other out, driving collective judgement closer to the truth [31–34]. In quantitative and probabilistic judgements, groups of diverse individuals often perform as well as or better than even the best-credentialed single expert [35, 36].

In the context of peer review, diverse experiences and worldviews may provide different perspectives and judgements on a manuscript [1, 37, 38]. This is particularly important because perceptions about the quality of new research are inherently subjective and socially constructed [39–41]. Perhaps for these reasons, reviewers have been shown to favour manuscripts on topics familiar to them [42, 43], and to undervalue manuscripts reporting results that contradict their prior beliefs [44] or written by members of traditionally underrepresented groups in academia [45, 46].

Editors will typically attempt to recruit reviewers who collectively cover the knowledge space



INVESTIGATE

All experts privately answer elicitation questions and provide rationales for their judgements.

DISCUSS

- Experts are shown anonymous answers and a visual summary of other participants' responses.
- Experts engage in (facilitated) discussion, focused on exploring the underlying reasoning, *not* on consensus.

ESTIMATE

All experts provide a 2nd private answer to the elicitation questions, accompanied by rationales for any changes in their judgements.

AGGREGATE

- Aggregated estimates are calculated, with the option of individual weightings (e.g. based on expertise, prior performance).
- Experts can review and discuss individual and aggregate estimates and correct residual misunderstandings.

Fig. 1 The IDEA protocol for structured expert judgement elicitation (adapted from [20])

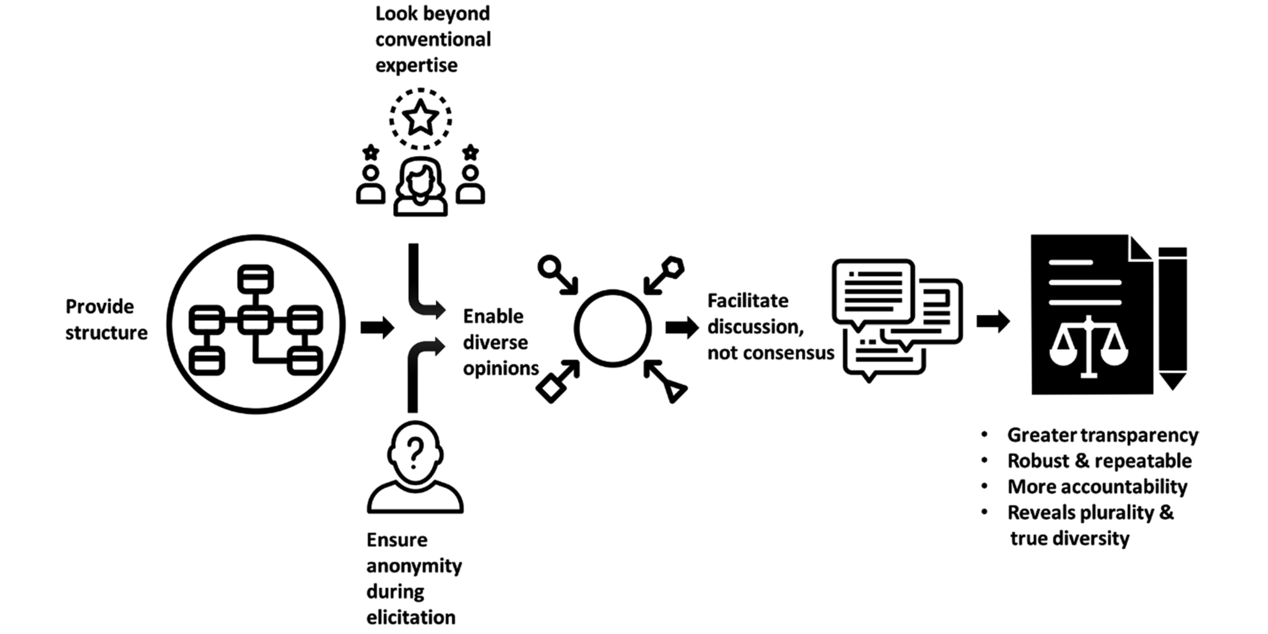

Look beyond conventional expertise
Provide structure
Enable diverse opinions
Facilitate discussion, not consensus
Ensure anonymity during elicitation
• Greater transparency
• Robust & repeatable
• More accountability
• Reveals plurality & true diversity

Fig. 2 A reimagined peer review process

corresponding to the topic of a paper, as would be advised in expert elicitations [47, 48]. This process, however, is often subject to bias. Recent research shows, for instance, that female reviewers have fewer opportunities to participate in peer review [49]. There is already ample discussion of the need to increase diversity and inclusion in peer review, but we argue that adopting more inclusive practices in peer review has benefits beyond achieving better representation. Demographic diversity (i.e., variation in age, gender, cultural background, life experience, education and specialization) can serve as a pragmatic proxy for the more crucial characteristic of cognitive diversity that underpins the wisdom of the crowd effect [50].

However, leveraging the collective insight of a diverse crowd of reviewers will not necessarily make the recruitment of peer reviewers more difficult. Most structured expert elicitation processes use between 3 and 7 experts. A recent survey of peer reviewing practices in 142 journals from 12 different disciplines and comprising 100,000 papers uncovered that on average each paper received 3.49 ± 1.45 (SD) reviews [51]. Accessing diverse "experts" may nevertheless be challenging, particularly in small, specialised fields, and require alternative approaches (e.g., recruiting graduate students, researchers in third sector organisations, government agencies, and industry), which leads us to our second recommendation.

Challenge conventional definitions of expertise

We tend to credit people with expertise based on their status, publications, affiliation with well-known institutions etc., yet such credentials have been shown to be unreliable criteria. Studies have shown mixed results for the association between traditional markers of expertise and indicators of judgement performance, such as overconfidence [52], reliability [53], calibration [54], and coherence [55]. Furthermore, selecting experts using conventional criteria can often bias the demographics of experts towards older individuals, and often males [28, 53]. To foster diversity, we must challenge our definition of expertise. Instead of setting our sights on a small population of narrowly defined "experts", our focus should be on engaging the wider scientific community, aiming to build skillsets in judging the quality of academic outputs through 'deliberate practice' [56, 57], which is a more relevant definition of expertise in this context.

Several large-scale projects have shown peer-reviewed research in medicine, psychology and economics often fails to replicate [58–63], raising fundamental questions about the scientific process, including peer reviewers' abilities to detect valid and reliable research. Yet recent studies have shown that, under the right conditions, groups of laypeople can accurately predict which research claims will replicate [64]. A computational analysis of the peer review process suggests "the accuracy of public reader-reviewers can surpass that of a small group of expert reviewers if the group of public reviewers is of sufficient size" [65]. Therefore, we argue that rather than relying on conventional expertise, the aggregate of judgements of groups of assessors with diverse knowledge, drawn from traditional and non-traditional reviewer pools, will result in more accurate judgements. Widening the potential reviewer pool may therefore convey benefits on review quality, in addition to addressing crucial ethical (i.e., increasing diversity and inclusion) and pragmatic concerns (i.e., reducing the burden on an already strained minority that generate the bulk of peer reviews [66]).

Provide structure

One aim of structuring elicitation processes is to quantify (but not eliminate) uncertainty. Structured expert elicitation protocols achieve this by standardising the response formats of quantitative elicitation questions, and IDEA, for instance, asks experts to provide bounds that give a measure of uncertainty. The procedure removes consensus pressure and associated biases by aggregating individual judgements mathematically rather than behaviourally. In peer review, this can be achieved by structuring judgements around a predefined set of questions about research quality, expressed in quantitative terms (some peer review rubrics do this already [67]).

Importantly, many structured expert elicitation protocols complement quantitative estimation in several ways, i.e. by collecting written feedback, facilitating discussion among experts, encouraging experts to correct misunderstandings, and by exploring diverse perspectives and counterfactual evidence without forcing consensus. A peer review process modelled after the IDEA protocol will generate both quantitative and qualitative data that will feed into editorial decision-making.

Nevertheless, the use of numerical scores may give an unwarranted impression of precision. Indeed, the extent to which different reviewers and/or the editors share conceptualisations or reference points for these scores is questionable and peer reviewers often are miscalibrated [68–70]. Notwithstanding these legitimate concerns, asking peer reviewers to provide numerical scores and interval judgements, in addition to narrative evaluations, may more readily highlight areas of disagreement and uncertainty. In discussion, expert groups may resolve some of their initial disagreements, while remaining disagreements may require the attention of the editor(s) and/or author(s).

Encourage and facilitate interaction

When faced with uncertainty, individuals – including experts – use heuristics, or cognitive shortcuts, that may lead to erroneous conclusions [76]. Group interaction increases the probability of identification and subsequent correction of errors [77] and can counter individual cognitive biases [78]. Interaction among experts also has a synergistic effect and may generate novel ideas and solutions that would not have been produced individually [71–75]. Nevertheless, the intellectual gain of group interaction is moderated by coordination costs [79], overconfidence [53, 80, 81] and deference to high-status or particularly vocal group members [82]. Expert elicitation protocols mitigate many of these risks by supporting experts to investigate and discuss the rationales that underpin individual judgements. The process aims to develop more comprehensive, unprejudiced understandings and create shared mental models [83]. By explicitly revealing private information, the process attempts to mitigate the pernicious effects of unshared information [84, 85] and prompts experts to imagine alternative realities, which reduces subjectivity and (over)confidence [86].

Applied to peer review, the interactive process we envision encourages reviewers to articulate their reasoning explicitly before sharing it with others, and subsequently to probe the rationales that underpin alternative judgments. It also promotes the resolution of conflicts due to misunderstandings, lack of relevant knowledge, or the application of idiosyncratic evaluation criteria [44, 87–90]. Importantly, it does not force consensus where true agreement does not exist. From the editor's point of view, having access to both outcome (individual reviewer decisions) and process data (interaction among reviewers) generates valuable insights into how reviewers' rationales withstand the scrutiny of their peers, distinguishes between trivial and fundamental differences of opinion, and ultimately enables a more informed and transparent editorial decision.

While interactions among reviewers are still relatively uncommon in journal peer review, interviews with editors and reviewers uncovered a practice of conferring with colleagues on how to review a manuscript and collective decision-making [91]. This suggests that both editors and peer reviewers may welcome opportunities to make the process more interactive.

Anonymise judgements

To further mitigate potential pernicious social influences that can undermine the wisdom of crowd [82], it is important to maintain some degree of independence, to enable individuals to express their true beliefs, without dominance or interpersonal conflict. Expert elicitation protocols like IDEA protect the anonymity of judgements [28] and encourage experts to evaluate each idea on its own merit, rather than on the status of its proponent, thus minimising some (although arguably not all) social biases.

Anonymity is a contested feature of traditional peer review protocols. Critics claim that it encourages laziness, idiosyncratic (or self-interested) reviews, responsibility and even abuses of power [91, 92]. Opening reviewers' judgments to the scrutiny of their peers and/or editors would mitigate some of these dangers or at least expose them before they are sent to authors and become enshrined in editorial decisions. What is more, reviewers' identities could also be revealed at the end of the elicitation process. So, by combining anonymous judgements with group interaction, the re-imagined peer review process outlined in this paper preserves the advantages of the traditional peer review system and mitigates the danger of unaccountable reviewers.

Outlook

Efforts are underway to support the implementation of more transparent peer-reviewing practices [93–95]. This commentary contributes to a wider conversation about establishing peer review processes grounded in evidence-based practice. Our suggestions are based on observations about how to identify experts, how to elicit judgments and how to manage their interactions that have been shown to reduce social influence effects and increase collective accuracy and calibration of judgements in other settings. To what extent similar effects can be achieved in peer review is an empirical question that remains unaddressed. In our previous work, expert elicitations were generally constrained to a simple two-round process, but in its application to peer review, this may have to be extended to involve any number of rounds to allow additional consideration of authors' responses upon resubmission. Moreover, there will be translational challenges in implementing structured expert elicitation protocols for the purposes of peer review, including the need to organise (at least for some manuscripts) synchronous discussion rounds between geographically (and time-zone) dispersed reviewers. Operationalising the above recommendations into everyday journal practices, across disciplines, will require some editorial bravery and careful experimentation.

Acknowledgements
We would like to thank Simine Vazire for helpful feedback on an earlier version of this paper. The icons used in Fig. 1 were obtained from the Noun Project (https://thenounproject.com/), created by the following artists: 'Sumit Saengthong', 'Icon 54', 'Brandon Shields', 'SBTS', and 'Nhor'.

Author contributions
Conceptualization: AM, AV, FF; Investigation, Methodology & Writing – original draft: AM, AV, MB, DGH, AH, VH, BW, MB, FF; Funding acquisition: FF; Project administration: AM. All authors read and approved the final manuscript.

Funding
This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. HR001118S0047. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Centre for Argument Technology, School of Science and Engineering (Computing), University of Dundee, Dundee, UK. [2]School of Communication and Arts, The University of Queensland, Brisbane, QLD, Australia. [3]MetaMelb Lab, University of Melbourne, Melbourne, VIC, Australia. [4]Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Melbourne, VIC, Australia. [5]Martin Conservation Decisions Lab, Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada. [6]Centre for Environmental Policy, Imperial College London, London, UK.

Received: 18 December 2021 Accepted: 24 March 2022
Published online: 05 April 2022

References

1. Fogg L, Fiske DW. Foretelling the judgments of reviewers and editors. Am Psychol. 1993;48(3):293–4.
2. Bornmann L, Daniel H-D. The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at angewandte chemie. Angew Chem Int Ed. 2008;47(38):7173–8.
3. Bornmann L, Mutz R, Daniel H-D. Row-column (RC) association model applied to grant peer review. Scientometrics. 2007;73(2):139–47.
4. Haffar S, Bazerbachi F, Murad MH. Peer review bias: a critical review. Mayo Clin Proc. 2019;94(4):670–6.
5. Baxt WG, Waeckerle JF, Berlin JA, Callaham ML. Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. Ann Emerg Med. 1998;32(3):310–7.
6. Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? J R Soc Med. 2008;101(10):507–14.
7. Anderson E. The need to review peer review: the regnerus scandal as a call to action. J Gay Lesbian Mental Health. 2013;17(3):337–51.
8. Francis JR. The credibility and legitimation of science: A loss of faith in the scientific narrative. Account Res. 1989;1(1):5–22.
9. Kravitz RL, Franks P, Feldman MD, Gerrity M, Byrne C, Tierney WM. Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care? PLoS ONE. 2010;5(4):e10072.
10. Helmer M, Schottdorf M, Neef A, Battaglia D. Gender bias in scholarly peer review. ELife. 2017;6:e21718.
11. Hamilton DG, Fraser H, Hoekstra R, Fidler F. Journal policies and editors' opinions on peer review. ELife. 2020;9:e62529.
12. Tennant JP, Ross-Hellauer T. The limitations to our understanding of peer review. Res Integr Peer Rev. 2020;5(1):6.
13. Walker R, da Silva P. Emerging trends in peer review—a survey. Front Neurosci. 2015;9:169.
14. Horbach SPJM, Halffman W. Journal peer review and editorial evaluation: cautious innovator or sleepy giant? Minerva. 2020;58(2):139–61.
15. Tennant JP, Dugan JM, Graziotin D, Jacques DC, Waldner F, Mietchen D, et al. A multi-disciplinary perspective on emergent and future innovations in peer review. F1000 Res. 2017;6:1151.
16. Peer review on trial. Nature. 2006;441(7094):668-.

17. Cooke R. Experts in uncertainty: opinion and subjective probability in science. Oxford: Oxford University Press; 1991.
18. Burgman MA. Trusting judgements: How to get the best out of experts. Cambridge: Cambridge University Press; 2016.
19. Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, et al. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. Psychol Sci. 2014;25(5):1106–15.
20. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: Eliciting experts' probabilities. Hoboken, NJ: Wiley; 2006.
21. Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc Natl Acad Sci. 2014;111(20):7176.
22. Hanea AM, McBride MF, Burgman MA, Wintle BC. Classical meets modern in the IDEA protocol for structured expert judgement. J Risk Res. 2018;21(4):417–33.
23. Hanea AM, McBride MF, Burgman MA, Wintle BC, Fidler F, Flander L, et al. I nvestigate D iscuss E stimate A ggregate for structured expert judgement. Int J Forecast. 2017;33(1):267–79.
24. Hemming V, Armstrong N, Burgman MA, Hanea AM. Improving expert forecasts in reliability: application and evidence for structured elicitation protocols. Qual Reliab Eng Int. 2020;36(2):623–41.
25. Hemming V, Hanea AM, Walshe T, Burgman MA. Weighting and aggregating expert ecological judgments. Ecol Appl. 2020;30(4):e02075.
26. Hemming V, Walshe TV, Hanea AM, Fidler F, Burgman MA. Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. PLoS ONE. 2018;13(6):e0198468.
27. Fraser H, Bush M, Wintle B, Mody F, Smith E, Hanea A, et al. Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessment for Trustworthy Science) [preprint]. OSF.io; 2021.
28. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. Methods Ecol Evol. 2018;9(1):169–80.
29. Wintle B, Mody F, Smith E, Hanea AM, Wilkinson DP, Hemming V, et al. Predicting and reasoning about replicability using structured groups. MetaArXiv Preprints. 2021.
30. Surowiecki J. The wisdom of the crowds: Doubleday; 2004.
31. Gigone D, Hastie R. The impact of information on small group choice. J Pers Soc Psychol. 1997;72(1):132–40.
32. Larrick RP, Soll JB. Intuitions about combining opinions: misappreciation of the averaging principle. Manage Sci. 2006;52(1):111–27.
33. Lorenz J, Rauhut H, Schweitzer F, Helbing D. How social influence can undermine the wisdom of crowd effect. Proc Natl Acad Sci. 2011;108(22):9020.
34. Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. Manage Sci. 2015;61(2):267–80.
35. Davis-Stober CP, Budescu DV, Dana J, Broomell SB. When is a crowd wise? Decision. 2014;1(2):79–101.
36. Tetlock PE, Gardner D. Superforecasting: The art and science of prediction: Random House; 2016.
37. Cole S. Making Science: Between Nature and Society. Cambridge: Harvard University Press; 1992.
38. Weller AC. Editorial peer review: Its strenghts and weaknesses. Medford, NJ: Information Today 2001.
39. Bedeian A. Peer review and the social construction of knowledge in the management discipline. Aca Manag Learn Educ. 2004;3:198–216.
40. Douglas H. Inductive risk and values in science. Phil Sci. 2000;67(4):559–79.
41. Longino H. How values can be good for science. Sci Values Obj. 2004;1:127–42.
42. Dondio P, Casnici N, Grimaldo F, Gilbert N, Squazzoni F. The, "invisible hand" of peer review: the implications of author-referee networks on peer review in a scholarly journal. J Informet. 2019;13(2):708–16.
43. Porter AL, Rossini FA. Peer review of interdisciplinary research proposals. Sci Technol Human Values. 1985;10(3):33–8.
44. Mahoney MJ. Publication prejudices: an experimental study of confirmatory bias in the peer review system. Cogn Ther Res. 1977;1(2):161–75.
45. Silbiger NJ, Stubler AD. Unprofessional peer reviews disproportionately harm underrepresented groups in STEM. Peer J. 2019;7:e824.
46. Lee CJ. Revisiting Current Causes of Women's Underrepresentation in Science. In: Michael Brownstein JS, editor. Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology: Oxford University Press; 2016.
47. Keeney RL, Winterfeldt DV. Eliciting probabilities from experts in complex technical problems. IEEE Trans Eng Manag. 1991;38(3):191–201.
48. Cooke RM, Goossens LHJ. Procedures guide for structural expert judgement in accident consequence modelling (invited paper). Radiat Prot Dosimetry. 2000;90(3):303–9.
49. Lerback J, Hanson B. Journals invite too few women to referee. Nature. 2017;541(7638):455–7.
50. Page SE. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies: Princeton University Press; 2007.
51. Raoult V. How Many Papers Should Scientists Be Reviewing? An Analysis Using Verified Peer Review Reports. Publications. 2020;8(1).
52. McKenzie CRM, Liersch MJ, Yaniv I. Overconfidence in interval estimates: What does expertise buy you? Organ Behav Hum Decis Process. 2008;107(2):179–91.
53. Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, et al. Expert Status and Performance. PLOS ONE. 2011;6(7):e22998.
54. Christensen-Szalanski JJ, Bushyhead JB. Physician's use of probabilistic information in a real clinical setting. J Exp Psychol Hum Percept Perform. 1981;7(4):928–35.
55. Mandel DR, Karvetski CW, Dhami MK. Boosting intelligence analysts' judgment accuracy: What works, what fails? Judgm Decis Mak. 2018;13(6):607–21.
56. Larkin JH, McDermott J, Simon DP, Simon HA. Models of competence in solving physics problems. Cogn Sci. 1980;4(4):317–45.
57. Ericsson K, Krampe R, Tesch-Roemer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev. 1993;100:363–406.
58. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483(7391):531–3.
59. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability. Soc Psychol. 2014;45(3):142–52.
60. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many labs 2: investigating variation in replicability across samples and settings. Adv Methods Pract Psychol Sci. 2018;1(4):443–90.
61. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. eLife. 2014;3:e04333.
62. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018;2(9):637–44.
63. Camerer Colin F, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016;351(6280):1433–6.
64. Hoogeveen S, Sarafoglou A, Wagenmakers E-J. Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully. Adv Methods Pract Psychol Sci. 2020;3(3):267–85.
65. Herron DM. Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. Surg Endosc. 2012;26(8):2275–80.
66. Publons. Global State of Peer Review. 2018.
67. Onitilo AA, Engel JM, Salzman-Scott SA, Stankowski RV, Doi SAR. A Core-Item Reviewer Evaluation (CoRE) System for Manuscript Peer Review. Account Res. 2014;21(2):109–21.
68. Siegelman SS. Assassins and zealots: variations in peer review. Spec Rep Radiol. 1991;178(3):637–42.
69. Ragone A, Mirylenka K, Casati F, Marchese M. On peer review in computer science: analysis of its effectiveness and suggestions for improvement. Scientometrics. 2013;97(2):317–56.
70. Shah NB. An overview of challenges, experiments, and computational solutions in peer review (extended version). Communications of the ACM; 2022.
71. Moshman M, Geil D. Collaborative reasoning: evidence for collective rationality. Think Reason. 1998;4(3):231–48.
72. Laughlin PR, Ellis AL. Demonstrability and social combination processes on mathematical intellective tasks. J Exp Soc Psychol. 1986;22(3):177–89.
73. Laughlin PR, Bonner BL, Miner AG. Groups perform better than the best individuals on Letters-to-Numbers problems. Organ Behav Hum Decis Process. 2002;88(2):605–20.
74. Woolley Anita W, Chabris Christopher F, Pentland A, Hashmi N, Malone TW. Evidence for a collective intelligence factor in the performance of human groups. Science. 2010;330(6004):686–8.

75. Mercier H, Trouche E, Yama H, Heintz C, Girotto V. Experts and laymen grossly underestimate the benefits of argumentation for reasoning. Think Reason. 2015;21(3):341–55.
76. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185(4157):1124–31.
77. Ziller RC. Group size: a determinant of the quality and stability of group decisions. Sociometry. 1957;20(2):165–73.
78. Schirrmeister E, Göhring A-L, Warnke P. Psychological biases and heuristics in the context of foresight and scenario processes. Fut Foresight Sci. 2020;2(2):e31.
79. Bates TC, Gupta S. Smart groups of smart people: evidence for IQ as the origin of collective intelligence in the performance of human groups. Intelligence. 2017;60:46–56.
80. Speirs-Bridge A, Fidler F, McBride M, Flander L, Cumming G, Burgman M. Reducing overconfidence in the interval judgments of experts. Risk Anal. 2009;30(3):512–23.
81. Hinsz VB, Tindale RS, Vollrath DA. The emerging conceptualization of groups as information processors. Psychol Bull. 1997;121(1):43–64.
82. Van De Ven A, Delbecq AL. Nominal versus interacting group processes for committee decision-making effectiveness. Acad Manag J. 1971;14(2):203–12.
83. Maciejovsky B, Sutter M, Budescu DV, Bernau P. Teams make you smarter: how exposure to teams improves individual decisions in probability and reasoning tasks. Manage Sci. 2013;59(6):1255–70.
84. Riedl C, Woolley AW. Teams vs. crowds: a field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance. Acad Manag Disc. 2016;3(4):382–403.
85. Stasser G, Titus W. Pooling of unshared information in group decision making: biased information sampling during discussion. J Pers Soc Psychol. 1985;48(6):1467–78.
86. Griffin D, Tversky A. The weighing of evidence and the determinants of confidence. Cogn Psychol. 1992;24(3):411–35.
87. Hojat M, Gonnella JS, Caelleigh AS. Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. Adv Health Sci Educ. 2003;8(1):75–96.
88. Church K. Reviewing the reviewers. Comput Linguist. 2006;31(4):4.
89. Smith R. Peer review: a flawed process at the heart of science and journals. J R Soc Med. 2006;99(4):178–82.
90. Lee CJ. Commensuration bias in peer review. Philos Sci. 2015;82(5):1272–83.
91. Lipworth WL, Kerridge IH, Carter SM, Little M. Journal peer review in context: A qualitative study of the social and subjective dimensions of manuscript review in biomedical publishing. Soc Sci Med. 2011;72(7):1056–63.
92. Goetz A. Open Science Collaboration. 2014. http://osc.centerforopenscience.org/2014/10/22/reexamining-reviewer-anonymity/.
93. Squazzoni F, Brezis E, Marušić A. Scientometrics of peer review. Scientometrics. 2017;113(1):501–2.
94. Bravo G, Grimaldo F, López-Iñesta E, Mehmani B, Squazzoni F. The effect of publishing peer review reports on referee behavior in five scholarly journals. Nat Commun. 2019;10(1):322.
95. Yan V. ReimagineReview News. 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.