The Influence of Disciplinary Background on Peer Reviewers' Evaluations of Engineering Education Journal Manuscripts

Abstract

This is the first of a series of studies that explore the relationship between disciplinary background and the weighting of various elements of a manuscript in peer reviewers' determination of publication recommendations. Research questions include: (1) To what extent are tacit criteria for determining quality or value of EER manuscripts influenced by reviewers' varied disciplinary backgrounds and levels of expertise? and (2) To what extent does mentored peer review professional development influence reviewers' EER manuscript evaluations? Data were collected from 27 mentors and mentees in a peer review professional development program. Participants reviewed the same two manuscripts, using a form to identify strengths, weaknesses, and recommendations. Responses were coded by two researchers (70% IRR). Our findings suggest that disciplinary background influences reviewers' evaluation of EER manuscripts. We also found evidence that professional development can improve reviewers' understanding of EER disciplinary conventions. Deeper understanding of the epistemological basis for manuscript reviews may reveal ways to strengthen professional preparation in engineering education as well as other disciplines.

Introduction

Philosophers, beginning with Plato [1], asked questions similar to those we explore here about how we know what we know and how knowledge is related to action. In the 20th century, the topic became relevant to psychologists and neuroscientists whose methods and tools enabled them to better study the workings of the mind and brain. Historically, researchers conceived of knowledge from a positivist perspective, fixed like a conventional computer program that is activated to inform or guide problem-solving using predefined code. Since the early 21st century, this view has been increasingly challenged, with theorists and practitioners both arguing that professional knowledge is transactional, socially constructed, and essentially functional, adapting to unique situations that professionals in particular face—analogous to a machine learning algorithm that is continually updated through experience. These discussions have infiltrated professional education, challenging conventional practices in higher education about how to prepare students to be teachers, architects, medical doctors, and engineers [2], [3].

Yet research on the epistemology of researchers is limited. Preparation of engineering education research (EER) professionals is a particularly interesting case, because engineering education is, like many of the social sciences, interdisciplinary. Some EER professionals were prepared in engineering education programs. Others were trained as engineers with no previous expertise in education research, but whose professional practice and intellectual interests motivated them to explore the teaching and learning of engineering. Others migrated into EER from social science disciplines, having no previous training in engineering [4].

All the paths that professionals have taken who study EER converge in manuscript review. Reviewers apply their professional knowledge in assessing the quality of a manuscript and the potential of a study to advance academic discourse and EER practice. Manuscript review

is a discussion (sometimes a negotiation) between EER professionals—reviewers, editors, and authors—about effective and robust EER practice. At the same time, manuscript review has a weighty, gate-keeping function. The decisions made about publications can have lasting effects on individuals, journals, and the profession itself. Even so, the basis by which manuscripts are evaluated is little known or understood. To what extent does one's disciplinary background influence one's assessments of the quality of a manuscript? Are tacit understandings about what constitutes value and quality of a study commonly shared among EER professionals? If they are shared, by what means do these shared understandings and professional values come to be so? If they are not shared, what, if anything, links the EER community? What characterizes EER as an intellectual community?

Literature Review

> Peer review clearly constitutes a social epistemic feature of the production and dissemination of scientific knowledge. It relies on members of knowledge communities to serve as gatekeepers in the funding and propagation of research. It calls on shared norms cultivated by the community. And it relies on institutions such as journal editorial boards, conference organizers, and grant agencies to articulate and enforce such norms [5, p. 868].

Not surprisingly, perhaps, researchers who have studied peer review typically focus on issues of reliability or convergence in the assessment of reviewers' ratings. The premise underlying these studies is that a manuscript has an inherent quality that can be assessed against the standards and conventions of an intellectual discipline, as long as reviewers are not corrupted by biases and/or inattentiveness during the review process [6], [7]. Several studies have explored bias in peer review [8]–[12]. Typical findings include low correlations between reviewers, bias in single-anonymous reviews (in which reviewers know the author's identity) that favor eminent researchers, and biases that favor prestigious institutions.

Most prior studies focus on the convergence of reviewers' publication recommendations (i.e., accept, minor revision, major revision, reject) or justification of their recommendations to the editor. Scant research has been conducted on the bases by which reviewers formulate those recommendations. In 1990, Fiske and Fog coded reviews of 153 initial submissions to American Psychological Association journals [13]. The 153 papers generated 402 reviews. The authors coded 3,477 weaknesses articulated in reviewer recommendations (strengths were not coded). The authors looked at the proportion of comments related to the conceptualization of the study, design, method, analysis, interpretations and conclusions, and presentation (quality of expression). Two-thirds of comments overall were related to the Planning & Execution of the study, and one-the third to the Presentation. Twenty percent of weaknesses were attributed to Conceptualization, 11% to Design, 12% to Procedures, 7% to Measurement, 22% to Analyses & Results, 16% to Interpretations/Conclusions, 9% to Editing/Writing, and 3% to "General." As with other studies of peer review, inter-class correlations of publication recommendations were low ($\bar{x}_r$=.20). That authors found minimal consensus across reviewers does not necessarily mean the reviewers disagreed. Fiske and Fog note, anecdotally, that the criticisms they coded all seemed appropriate. They also found very few disagreements across reviewers about specific issues in the paper. We surmise that variability in the reviewers' recommendations resulted from

individuals weighting specific strengths and weaknesses differently. Most recommendations to editors are brief summaries, which likely do not contain all of the tacit criteria reviewers used to evaluate the manuscript. In many reviews, only the most prominent features of a manuscript—negative and positive—are likely to be mentioned. This raises the question, what factors influence how reviewers weigh various factors that result in their recommendation on whether to publish a manuscript?

In this study, we explore the relationship, if any, between disciplinary background and the weighting of various elements of a manuscript since one's understanding of effective practice is likely shaped by one's understanding of the questions, methods, and evidence one brings to bear to discipline-related problems [14]. The following research questions guided this study:

1. To what extent are tacit criteria for determining quality or value of EER manuscripts influenced by reviewers' varied disciplinary backgrounds and levels of expertise?
2. To what extent does mentored peer review professional development influence reviewers' EER manuscript evaluations?

Methods

Overview of study

Data were collected from participants in a peer reviewer training program [15]. Participants in the program (n= 27; 9 mentors and 18 mentees) were formed into nine triads that each completed three peer reviews of manuscripts submitted to an EER journal. Participants were also invited to participate in think-aloud protocols, exit interviews, and focus groups, complete an exit survey and fill out Structured Peer Review (SPR) Forms before (Pre-SPR) and after (Post-SPR) participation in the program. All recruitment and data collection procedures followed approved IRB protocols. This paper focuses on responses to the Pre- and Post-SPRs.

Participant Selection and Background

The 18 mentees in the cohort were selected through a competitive, online application process that collected contact information, demographic information, and professional background, specifically about their Ph.D. concentration and year of degree, current position, relevant EER experience (e.g., publications, presentations, and reviewing history), confidence reviewing EER manuscripts, and the number of EER colleagues with whom they regularly interact. Special consideration was given to individuals deemed "lone wolves" who were not well-connected to an EER network and diverse participants who may not have been previously connected to the EER community [16]. Mentors were invited to participate based on their experience in EER, recommendations from journal editors and colleagues, and their desire to help advance EER through peer review. Mentors provided similar background information as mentees, with additional questions related to mentoring experience. In total, mentors and mentees represented 24 universities in six countries. Participants' experience levels included graduate students, postdoctoral researchers, and faculty. Mentees had varied backgrounds in social sciences, engineering, and engineering education.

SPR Data Collection and Code Development

Triads (one mentor and two mentees) were formed based on participants' time zones and areas of expertise. Each triad completed reviews of three articles submitted to an EER journal (Figure

1). Participants were asked to complete an SPR prior to the triad's first meeting (Pre-SPR), for each manuscript they reviewed as a triad (SPRs 1, 2, and 3), and after their final triad review was submitted (Post-SPR). In all, 26 out of the 27 program participants consented to be part of the research study, and their responses on the Pre- and Post-SPR are the source of data for this paper. The SPR is an online questionnaire that asks participants to describe the five most notable strengths and weaknesses of a manuscript, recommend a decision to the editor (accept, minor revision, major revision, or reject), and make a 200-word justification of their recommendation (Figure 2). Participants were instructed to fill out their SPRs individually and use them as a starting point in their discussions with their triads [15].

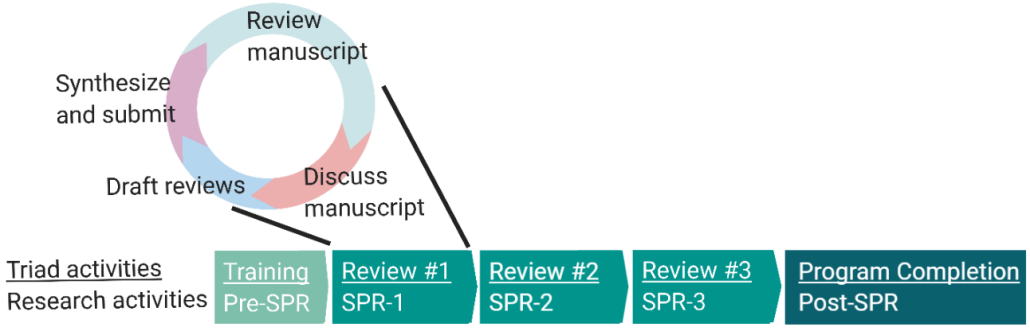
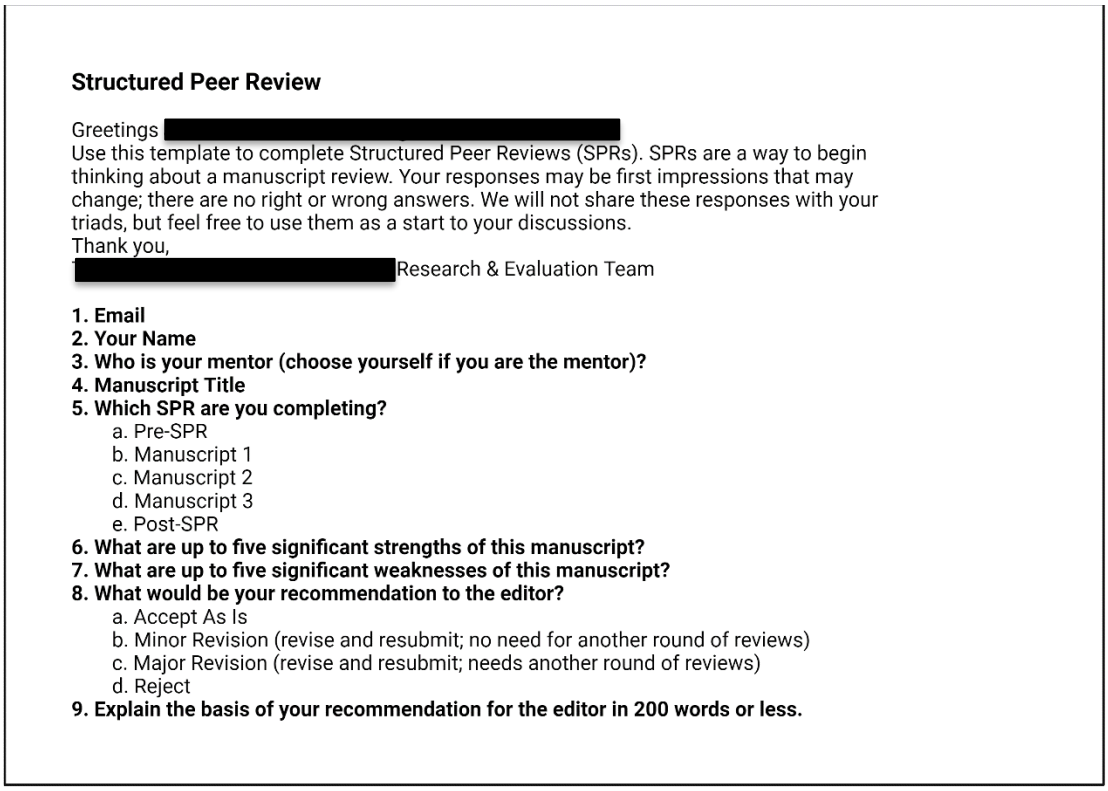


Figure 1: Activities completed as part of the peer reviewer training program. Each triad collaborated to complete three manuscript reviews. Each participant was asked to individually complete a Pre-SPR at the beginning of the program, for each of the three manuscripts they reviewed as a triad, and a Post-SPR at the end of the program.[15]

The Pre- and Post-SPR results were based on ~1500-word manuscripts that had been submitted to a special edition of a peer-reviewed EER journal and were used with permission from the authors for our research purposes. For both the Pre- and the Post- articles, the associate editor recommended "major revision" after receiving recommendations of both "major revision" or "reject" by journal reviewers. In all, triads completed five SPRs: a Pre-SPR completed by all mentors and mentees on the same manuscript, three SPRs on triad-specific manuscripts, and a Post-SPR completed by all participants on a different manuscript than the Pre.

SPR codes are central to the analysis. The first cohort of the program responded to the Pre-SPR manuscript using open-ended responses pertaining to manuscript strengths, weaknesses, and recommendations. Using a grounded theory approach [17], project researchers used open coding to identify responses that described similar ideas. These were reviewed and revised iteratively and then further refined through axial coding. Axial coding resulted in identifying six groups of codes, which we define as *themes*: *Context, Methods, Results, Discussion, Mechanics and Structure,* and *EER Relevance.* Within each theme, codes were organized as strengths (positive attributes) and weaknesses (negative attributes) (Figure 3). Once codes were finalized, they were inserted into the SPR form as checkbox lists that respondents could select from within strengths and weaknesses. The 200-word justification of the recommendation to the editor was kept as an open-ended response in the Post-SPR. To ensure that codes were not simply an artifact of a single manuscript, the process was repeated for a subsequent cohort reviewing a different manuscript. Few revisions were made, and the codes have been validated during repeated trials.



Structured Peer Review

Greetings [redacted]
Use this template to complete Structured Peer Reviews (SPRs). SPRs are a way to begin thinking about a manuscript review. Your responses may be first impressions that may change; there are no right or wrong answers. We will not share these responses with your triads, but feel free to use them as a start to your discussions.
Thank you,
[redacted]Research & Evaluation Team

1. Email
2. Your Name
3. Who is your mentor (choose yourself if you are the mentor)?
4. Manuscript Title
5. Which SPR are you completing?
a. Pre-SPR
b. Manuscript 1
c. Manuscript 2
d. Manuscript 3
e. Post-SPR
6. What are up to five significant strengths of this manuscript?
7. What are up to five significant weaknesses of this manuscript?
8. What would be your recommendation to the editor?
a. Accept As Is
b. Minor Revision (revise and resubmit; no need for another round of reviews)
c. Major Revision (revise and resubmit; needs another round of reviews)
d. Reject
9. Explain the basis of your recommendation for the editor in 200 words or less.

Figure 2: The open-ended Structured Peer Review (SPR) form distributed to participants prior to their first triad meeting was used to determine what criteria participants used to evaluate manuscripts when conducting their reviews and making a recommendation to the editor.

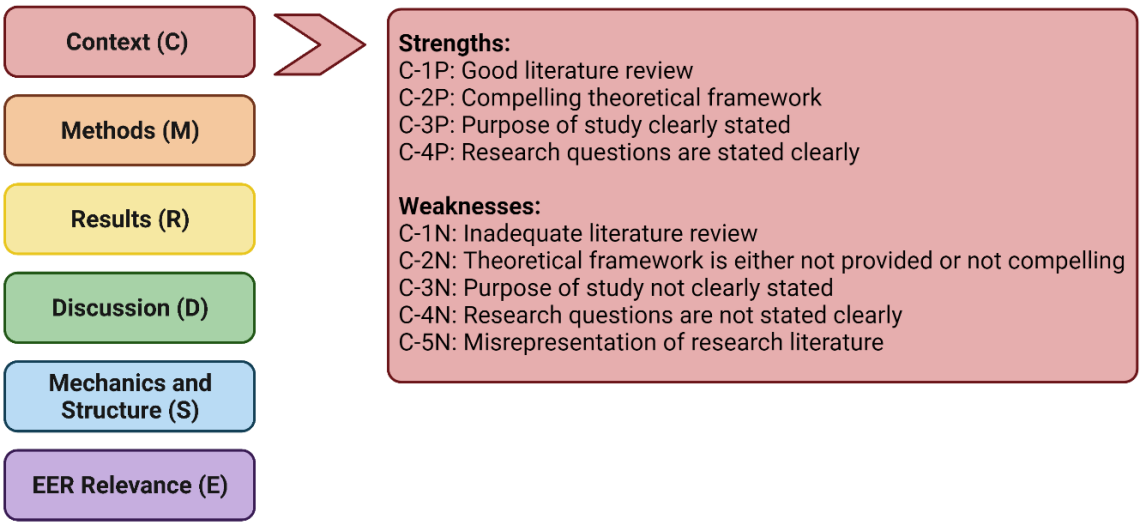


Figure 3: The six themes used for characterizing responses on the Pre- and Post-Structured Peer Review (SPR) forms: *Context, Methods, Results, Discussion, Mechanics and Structure*, and *EER Relevance*. Each of these themes had multiple codes organized as strengths (positive attributes) and weaknesses (negative attributes). For example, *Context* had four strengths (P for positives) and five weaknesses (N for negatives).

After each participant completed all triad manuscript reviews, they were sent the Post-SPR manuscript to review. Participants identified strengths and weaknesses from the checkbox lists, then wrote 200-word, open-ended justifications to the editor. A project researcher parsed recommendations into segments (phrases or sentences). Each segment was coded independently by two researchers using the SPR codes. Interrater reliability (IRR) was calculated as the number of segments that reflected agreement between the two raters divided by total segments. Although some 200-word responses included the same code more than once, any one code was only counted once per response. IRR between the two coders was 70%. Although no standards exist for inter-rater reliability for qualitative data, a reliability rating of r= 0.70 on open-coding of phenomenological data can be considered an acceptable cut-point [18], [19].

Data Analysis

After IRR was determined, analyses were conducted only on segments upon which both coders agreed. For analysis of these data, we report results for codes used by at least 50% of reviewers, which we define as "convergence," in response to the three SPR questions (strengths, weaknesses, and justification of recommendation to the editor). It is important to note that some codes had 50% convergence in strengths *and* the 200-word justification or in weaknesses *and* the 200-word justification. Our research team made the assumption that codes reflected in the strengths/weaknesses and justification were more important to reviewers than codes that were only identified within strengths/weaknesses and not mentioned in the justification. In each figure below, the codes referenced in strengths or weaknesses *and* in the justification are denoted with an asterisk (for example, C:4N* indicates *Context*, code 4 negative was cited by at least 50% of participants in both weaknesses *and* the recommendation justification).

Results

The Pre- and Post-SPR response rates were both 96% (*n*=25 of 26 participants). Results are reported for all participants (Figure 4), by role in the program (mentors—(MN), and mentees—(ME), Figure 5), by Ph.D. discipline concentration (engineers—(E), engineering educators—(EE), and social scientists—(SS), Figure 6), and by recommendation to the editor (major revision—(MR) and reject—(R), Figure 7).

'Context' showed convergence in participants' responses across the Pre- and Post-SPR

Before analyzing by discipline, we first determined which criteria participants converged on in their reviews. All codes used by at least 50% of the participants as strengths, weaknesses, or justifications of their recommendations to editors are shown in Figure 4. Codes used by at least 50% of the participants in the Pre-SPR and Post-SPR were C-2N (*Context: Theoretical framework not provided or not compelling*) and E-3P (*EER Relevance*: *Relevant to EER and/or timely*). Across the Pre-SPR and Post-SPR, the codes that had the most convergence were related to *Context*. Code C-3N was used by at least 50% of participants for the Pre-SPR, and codes C-1N and C-4N* were used by at least 50% of participants for the Post-SPR. These codes for *Context* are outlined in Figure 3 and relate to problem framing (i.e., literature review, theoretical framework, purpose, and research questions). Additionally, Methods code M-5N was cited as a weakness by at least 50% of participants in their responses to the Pre-SPR.

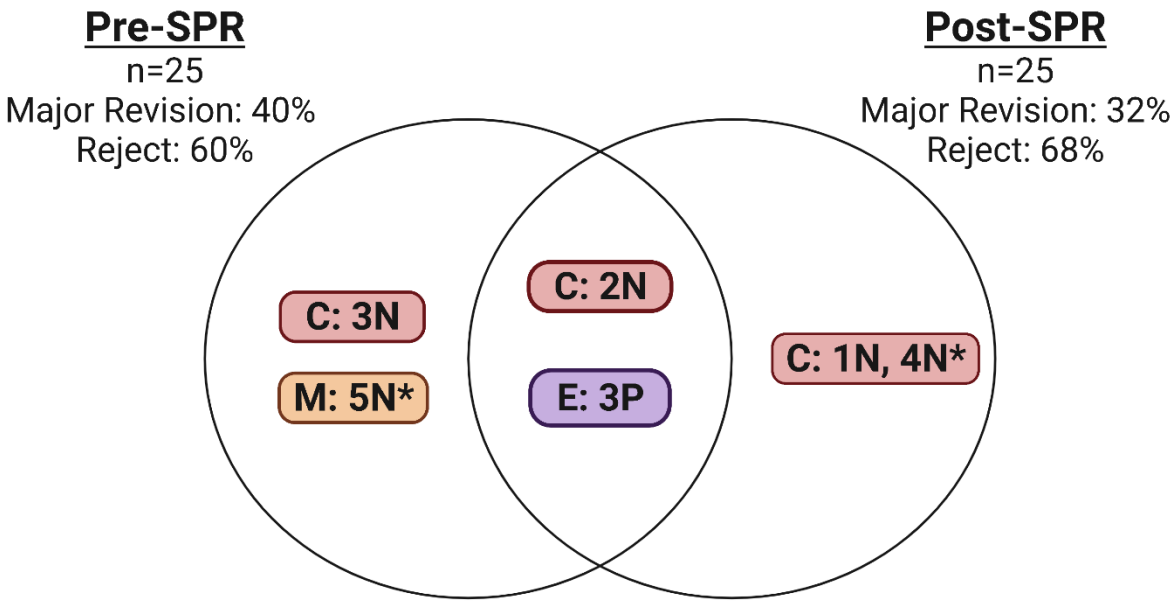


Figure 4: Codes used by at least 50% of participants in their responses to questions about strengths, weaknesses, and 200-word justification of their recommendation to the editor in the Pre-SPR and Post-SPR. Any code used at least 50% of the time by participants in either the strengths and recommendation to the editor or weaknesses and recommendation to the editor is denoted with an asterisk (*).

From Pre- to Post-, mentee codes became more aligned with those of mentors
We found that mentors' Pre-SPR reviews evidenced more convergence than the mentees' reviews (Figure 5). There was only one code (E-3P) that both mentors and mentees used at least 50% of the time in their responses on the Pre-SPR; however, only mentors used this code in both their responses to the weaknesses and 200-word justification. Mentors cited five additional codes in common (C-1N, C-2N, C 3N*, M-5N*, and S-2P). Mentees had no additional shared alignment of codes in the Pre-SPR.

At the end of the peer mentoring program, mentors and mentees were more aligned in their reviews of the Post-SPR manuscript. Both mentors and mentees used three codes (C-2N, C-4N, and E-3P) in at least 50% of their Post-SPR responses. Additionally, mentors and mentees converged on their use of C-4N in both the weaknesses and the 200-word justification. Mentors had one additional code (C-3N), and mentees had two additional codes (C-1N and M-1P) in common on at least 50% of their responses to the Post-SPR.

On Pre-SPRs, mentees were more likely than mentors to recommend the manuscript be rejected (71% of mentees compared to 27% of mentors). Recommendation decisions also became more aligned in the Post-SPR, with 69% of mentees and 67% of mentors recommending rejecting the manuscript.

Disciplinary differences were identified in the Pre-SPR and Post-SPR
We explored whether differences emerged based on participants' disciplinary backgrounds. Participants' responses were grouped based on their Ph.D. discipline: social sciences, engineering, and engineering education. This analysis, shown in Figure 6, only includes data

from mentee participants because the research team assumed that mentors were more likely to have internalized EER conventions and criteria, regardless of their disciplinary background.

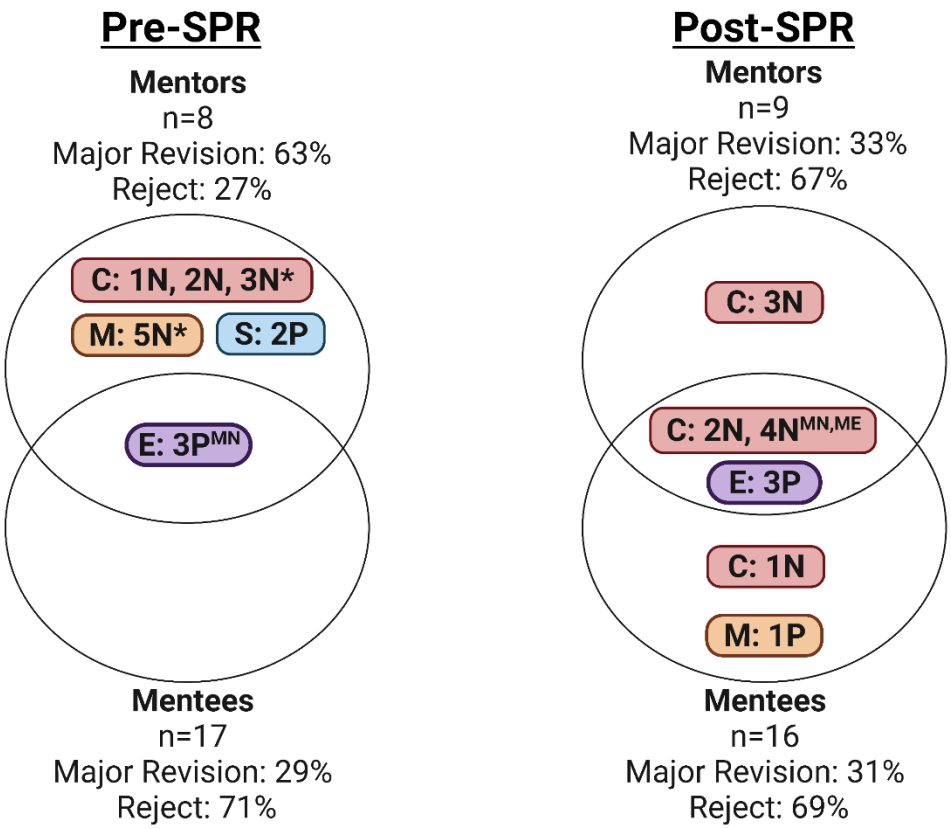


Figure 5: Mentors (MN) and mentees (ME) became more aligned in their reviews on the Post-SPR compared to the Pre-SPR. Code E-3P^{MN} was the only code used at least 50% of the time by both mentors and mentees in their Pre-SPR responses. In Post-SPR responses, codes C-2N, C-4N$^{MN, ME}$, and E-3P were all used at least 50% of the time by both mentors and mentees.

There were two codes (C-2N and E-3P) that at least 50% of mentees, regardless of discipline, used in their responses to the Pre-SPR. Those with social science backgrounds had the most convergence in their reviews, with four additional codes (C-3N*, C-4N, M-6N, and S-2N) used by at least 50% of participants. We identified one code (M-5N^{SS}) that at least 50% of those with social science and engineering education backgrounds shared in their Pre-SPR responses. Additionally, engineering educator Ph.Ds. converged on code C-1N, and engineering Ph.Ds. converged on code S-3N.

In Post-SPRs, mentees were more aligned across disciplines. There were three codes (C-1N^{SS}, C 4N^{SS}, and M-2N) that all mentees used at least 50% of the time. Those with social science and engineering backgrounds had four additional shared codes (M-1P, M-5N^{SS}, S-2N, and E-3P). Engineering educators and engineers had one shared code (C-2N$^{EE, E}$).

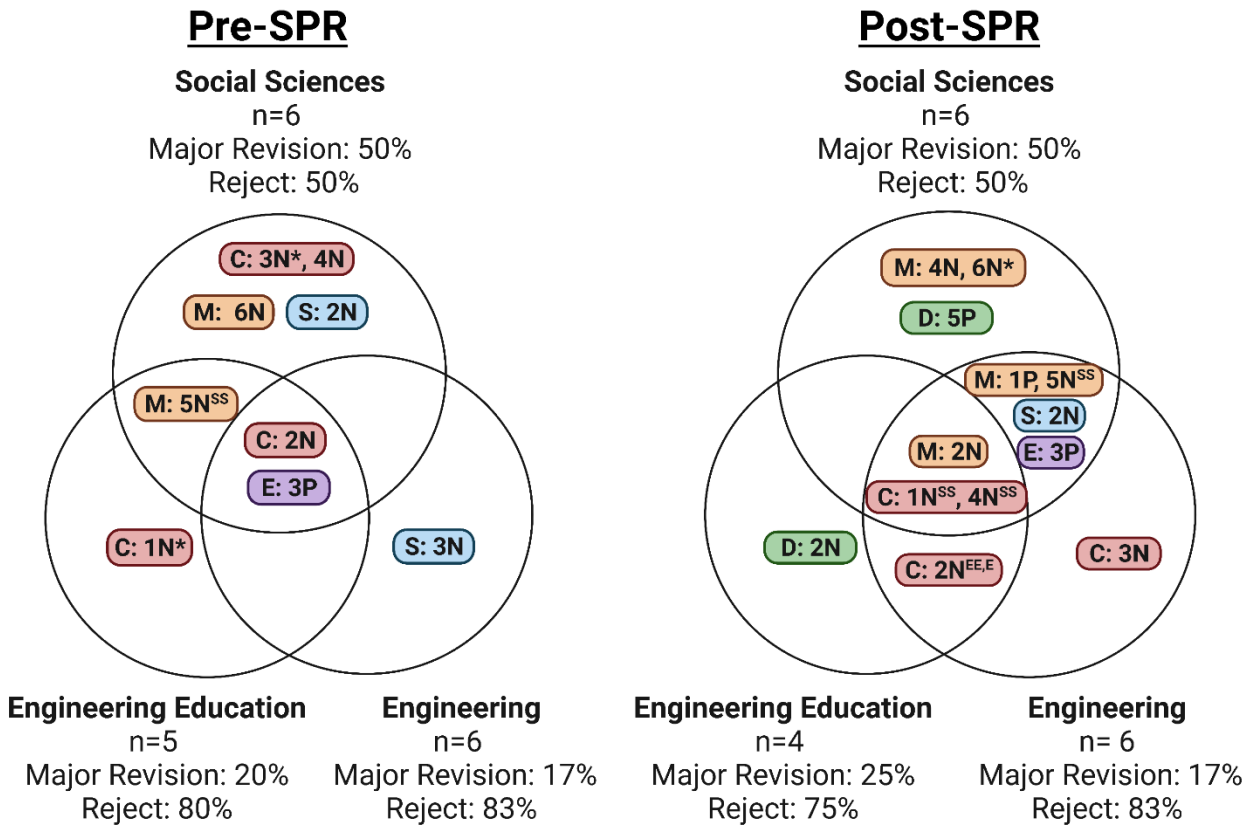


Figure 6: Disciplinary backgrounds include engineers (E), engineering educators (EE), and Social Scientists (SS). There was more alignment among and between disciplinary backgrounds in criteria used to evaluate reviews in the Post-SPR compared to the Pre-SPR. Codes C-2N and E-3P were the only codes used by at least 50% of the participants from all disciplines in the Pre-SPR responses, and one code, M-5N, was used by those in both social sciences and engineering education. In the Post-SPR responses, codes C-1N^{SS}, C-4N^{SS}, and M-2N were all used by at least 50% of the mentees from all disciplines; codes M-1P, M-5N^{SS}, S-2N, and E-3P were used by those in both social sciences and engineering; code C-2N$^{EE, E}$ was used by those from engineering and engineering education.

Participants who recommended 'major revision' converged more on positive codes than those who recommended 'reject' in the Post-SPR

Although mentees' Post-SPR codes reflected increased convergence across disciplines, their recommendation decisions did not. The ratio of "major revision" to "reject" recommendations percentages remained consistent for those from the social sciences (3:3) and engineering (1:5) for both the Pre-SPR and Post-SPR and only changed slightly for those with engineering education backgrounds (1:5 for the Pre-SPR and 1:4 for the Post-SPR). Note that there was also one less respondent for the Post-SPR; of the 23 participants who completed both a Pre-SPR and Post-SPR, only 5 (22%) used a different recommendation to the editor, even though it was two different manuscripts.

As Figure 7 illustrates, the criteria used to make recommendations were relatively consistent, suggesting that reviewers who recommend "major revision" or "reject" are likely weighing the same criteria differently. In the Pre-SPR, at least 50% of responses were coded as C-2N, M-5N^{MR}, and E-3P^{MR} for both "major revision" and "reject" recommendations to the editor (no

participants recommended "minor revision" or "accept"). Likewise, three codes (C-1N, C-4N$^{MR, R}$, and E-3P) were used for at least 50% of the participants who recommended "major revision" and "reject." One exception to this observation was the amount of alignment on the positive codes used in the Post-SPR compared to the Pre-SPR for those who recommended 'major revision.' In addition to E-3P, at least 50% of participants who recommended "major revision" in the Post-SPR agreed on three additional positive codes (D-5P, S-1P, and M-1P). However, these strengths are not well-aligned with their 200-word justification to the editor (note the lack of superscript on positive codes). This indicates that participants are likely using criteria that they do not always explicitly express in the 200-word justification of their recommendation to the editor.

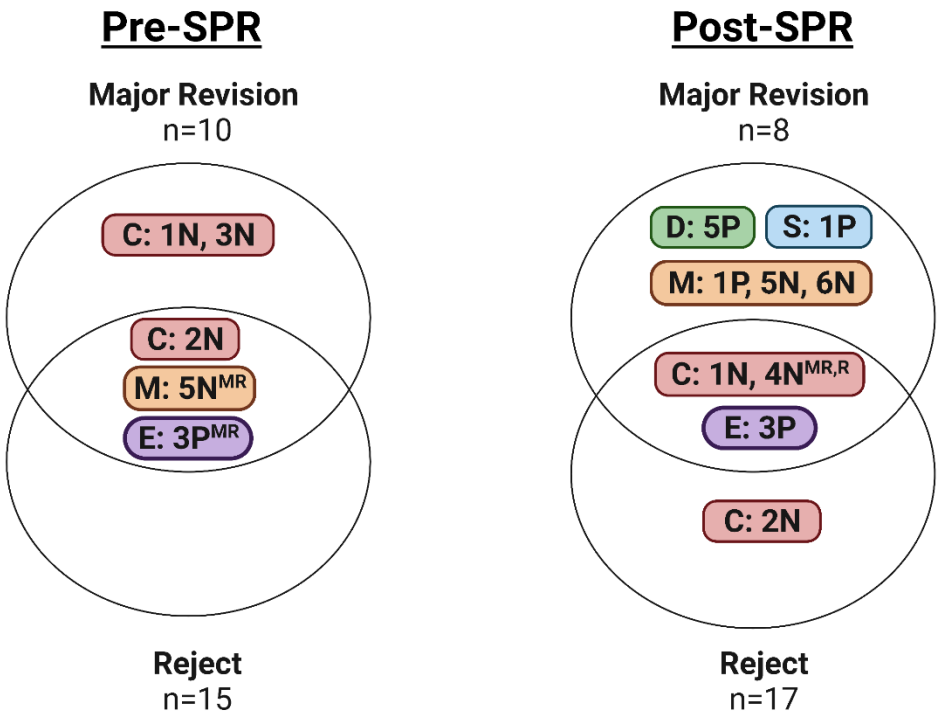


Figure 7: There was consistent code alignment for participants' recommendations of "major revision" or "reject." In their Pre-SPR responses, at least 50% of "reject" and "major revision" recommendations were coded as C-2N, M-5N^{MR}, and E-3P^{MR}. In their Post-SPR responses, at least 50% of "reject" and "major revision" recommendations were coded as C-1N, C-4N$^{MR, R}$, and E-3P. There was more alignment in positive codes (M-1P, D-5P, and S-1P) in responses for those who recommended "major revision" than "reject" in the Post-SPR compared to the Pre-SPR.

Limitations

A limitation of this study is the small sample (n=27), resulting in few cases within each subgroup analyzed. Future data collection will address this deficiency. Also, it is possible that increased alignment of codes among participants in the Post-SPR compared to the Pre-SPR could be due to the fact that codes were provided on the Post-SPR, whereas the Pre-SPR was open-ended. To address this issue, we used the Post- manuscript from cohort 1 as the Pre-manuscript for cohort 2 and made the SPR open-ended. Doing this enables us to assess the extent to which open coding aligns with check boxed coding. As of this writing, we have not

finished the analysis of cohort 2 data. Flipping Pre- and Post- manuscripts across cohorts also enables us to investigate the extent to which Pre- vs. Post-SPR codes reflect distinct characteristics of each manuscript as opposed to reviewers' predilections. Finally, the manuscripts used for the Pre- and Post- were each only ~1500 words, raising the possibility that manuscripts were not valid reflections of full research articles. Use of these abbreviated samples enabled the research team to collect a large number of reviews of the same manuscript. The project has also collected three SPRs of full-length manuscripts for each triad. Future analyses will compare our results from the full cohort reviews to those of the triads.

Discussion

Scholars in EER come from a wide variety of academic backgrounds. Many were trained as engineers, some as social scientists, and an increasing number as engineering educators. We hypothesized that those from different disciplines, especially those new to the EER field, may perceive the strengths and weaknesses of EER manuscripts differently based on the different training and experiences they have had in their respective fields. Specifically, we asked: To what extent are tacit criteria for determining quality or value of EER manuscripts influenced by reviewers' varied disciplinary backgrounds and levels of expertise? At the start of the program, we found that there was little convergence between the mentors and the mentees, and there was convergence among the mentors. Likewise, when the mentees were analyzed by discipline, there was little convergence among disciplines, but within the disciplines, especially those from the social sciences, there was the alignment of codes, providing evidence of disciplinary differences in manuscript evaluation.

Codes relating to *Context* had the most convergence within and across subgroup comparisons. This could potentially indicate that uniformly, regardless of background, criteria related to problem framing is a common factor. The peer-review study conducted by Fiske and Fog reported a similar finding [13].

Mentors, who are more experienced in EER, had six aligned codes applied to their responses to the Pre-SPR, but mentees only had one (which was also shared by mentors). When we analyzed the Pre-SPR by discipline, mentees across all disciplines converged on two applied codes, with those from engineering education and social science backgrounds converging on one additional applied code. Notably, Pre-SPRs from the social sciences had four other applied codes they converged on, providing potential evidence of disciplinary discrepancies in manuscript evaluation. As our results illustrate, these differences between groups became less pronounced after the peer mentoring program in their responses to the Post-SPR. This suggests, as we hypothesized, that as scholars gain more experience in EER, potential disciplinary differences become less pronounced.

We are in the early stages of our study, yet we find implications from the data so far. Most researchers receive little or no training in peer review. However, as increasing numbers of engineering education researchers are involved in peer review of journal and conference manuscripts, it is essential to consider the extent to which understanding of quality in EER research is shared. Notably, after completion of the program, there was greater convergence between mentors and mentees across disciplines. The triad model could promote this convergence. Although there was minimal interaction across triads, in the Post-SPR, we observed increased convergence between the mentors and mentees and among mentees from different disciplines. This suggests that there are epistemological foundations upon which EER

professionals evaluate manuscripts and that these conventions can become shared through peer mentoring.

One factor that did not appear to be influenced by the peer reviewer training program was the ratio of "major revision" to "reject" recommendations for the two different manuscripts. Only 5 of the 23 participants who completed both a Pre-SPR and Post-SPR changed from one recommendation to the other, even though the manuscripts were different. Additionally, the convergence of applied codes between those who recommended "major revision" and "reject" was similar in the Pre-SPR and Post-SPR. We wonder whether individual reviewers are predisposed to recommend "reject" over "major revision" or vice versa. The primary difference we observed between the codes applied to "major revision" and "reject" recommendations was the more frequent use of positive codes when asked about the strengths of the manuscript for those who recommended "major revision." Reviewers who recommended "major revision" converged on four positive codes, none of which were included in their justifications to the editor. This indicates that consideration of both strengths and weaknesses is important when studying the criteria peer reviewers use to evaluate manuscripts. The manuscripts used in this study only produced recommendations of "major revision" or "reject" from peer reviewers of the original manuscripts submitted for publication, which aligns with the recommendations from participants in our study. In future studies, we will use manuscripts expected to produce more diverse recommendations, which should improve our understanding of how positive criteria are considered by reviewers when evaluating manuscripts and if these criteria are impacted by background discipline.

Conclusion

This paper explored the aspects of EER manuscripts that peer reviewers notice and comment on in their reviews and recommendations to editors. Our data are unique in including recommendation decisions as well as justifications and strengths and weaknesses that might or might not have been included in publication justifications. These preliminary findings suggest that disciplinary background does play a role in the criteria that reviewers use to evaluate manuscripts. We also found evidence that peer review professional development can influence how reviewers understand EER research quality. This can enhance the preparation of novice peer reviewers, whether they are graduate students or seniors in the field making a transition to EER. In no way are we arguing for standardization in peer review. Assessment of "quality research" must be informed by one's professional knowledge and experience, and in this, we would expect variation. Yet, in a field as new as EER, discussions about the criteria about which we judge quality can promote enhanced understanding of the questions we ask and the methods we use to explore them. Deeper understanding of the epistemological basis for manuscript reviews may reveal ways to strengthen professional preparation in engineering education as well as in other disciplines.

Future research will involve additional participants from cohorts who continue to participate in the peer reviewer training program. We will continue to collect and analyze data from these new cohorts, which will provide a more robust sample size and a broader range of disciplinary backgrounds.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2037807, 2037788, and 2037797. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank the program participants for sharing their experiences and the mentors and coaches who support the program. The authors also thank the project advisory board members Rocio Chavela, Bev Watford, and Darryl Williams for their guidance on the project. Additional thanks to the reviewers who took the time to provide us with thorough feedback that helped shape the manuscript. All figures in the manuscript were generated using BioRender.

References

[1] "Epistemology - The history of epistemology | Britannica." https://www.britannica.com/topic/epistemology/The-history-of-epistemology (accessed Feb. 08, 2022).
[2] C. Coles, "Developing professional judgment," *J. Contin. Educ. Health Prof.*, vol. 22, no. 1, pp. 3–10, 2002, doi: 10.1002/chp.1340220102.
[3] D. A. Schön, *The Reflective Practitioner: How Professionals Think in Action*. London: Routledge, 2017. doi: 10.4324/9781315237473.
[4] L. C. Benson, K. Becker, M. M. Cooper, O. H. Griffin, and K. A. Smith, "Engineering Education: Departments, Degrees and Directions," *Eng. Educ.*, p. 7.
[5] C. J. Lee, "A Kuhnian Critique of Psychometric Research on Peer Review," *Philos. Sci.*, vol. 79, no. 5, pp. 859–870, Dec. 2012, doi: 10.1086/667841.
[6] R. K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press, 1979. Accessed: Feb. 08, 2022. [Online]. Available: https://press.uchicago.edu/ucp/books/book/chicago/S/bo28451565.html
[7] T. R. Tyler, "Psychological Perspectives on Legitimacy and Legitimation," *Annu. Rev. Psychol.*, vol. 57, no. 1, pp. 375–400, 2006, doi: 10.1146/annurev.psych.57.102904.190038.
[8] S. J. Ceci and D. P. Peters, "Peer Review: A Study of Reliability," *Change Mag. High. Learn.*, vol. 14, no. 6, pp. 44–48, Sep. 1982, doi: 10.1080/00091383.1982.10569910.
[9] J. R. Gilbert, E. S. Williams, and G. D. Lundberg, "Is There Gender Bias in JAMA's Peer Review Process?," *JAMA*, vol. 272, no. 2, pp. 139–142, Jul. 1994, doi: 10.1001/jama.1994.03520020065018.
[10] M. Hojat, J. S. Gonnella, and A. S. Caelleigh, "Impartial Judgment by the 'Gatekeepers' of Science: Fallibility and Accountability in the Peer Review Process," p. 22.
[11] L. Manchikanti, "Medical Journal Peer Review: Process and Bias," *Pain Physician*, vol. 18;1, no. 1;1, pp. E1–E14, Jan. 2015, doi: 10.36076/ppj/2015.18.E1.
[12] J. R. Cole and S. Cole, *Peer Review in the National Science Foundation: Phase II*. National Academy Press, 2101 Constitution Avenue, N, 1981.
[13] D. W. Fiske and L. F. Fogg, "But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments," *Am. Psychol.*, vol. 45, no. 5, pp. 591–598, 1990, doi: 10.1037/0003-066X.45.5.591.
[14] S. Toulmin, *Human Understanding: General introduction and pt. 1, The collective use and evolution of concepts*. 1972.

[15] L. Benson *et al.*, “Building Research Skills through Being a Peer Reviewer,” presented at the 2021 ASEE Virtual Annual Conference Content Access, Jul. 2021. Accessed: Feb. 08, 2022. [Online]. Available: https://peer.asee.org/building-research-skills-through-being-a-peer-reviewer
[16] D. Riley, J. Karlin, J. Pratt, and S. Quiles-Ramos, “Board # 127 : Building Social Infrastructure for Achieving Change at Scale,” in *2017 ASEE Annual Conference & Exposition Proceedings*, Columbus, Ohio, Jun. 2017, p. 27722. doi: 10.18260/1-2--27722.
[17] J. M. Corbin, *Basics of qualitative research: techniques and procedures for developing grounded theory*, Fourth edition. Thousand Oaks, California: SAGE, 2015.
[18] J. Marques and C. McCall, “The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study,” *Qual. Rep.*, vol. 10, no. 3, pp. 439–462, Sep. 2005, doi: 10.46743/2160-3715/2005.1837.
[19] M. B. Miles, A. M. Huberman, and J. Saldana, *Qualitative Data Analysis*. SAGE, 2014.