The beauty and the pain of peer review in Software Engineering

Júlia Condé Araújo
Colorado State University
Fort Collins, CO, USA
julia.condearaujo.colostate.edu

Marina Condé Araújo
Colorado State University
Fort Collins, CO, USA
marina.condearaujo.colostate.edu

Fabio Marcos de Abreu Santos
Colorado State University
Fort Collins, CO, USA
fabio.deabreusantos@colostate.edu

Bianca Trinkenreich
Colorado State University
Fort Collins, CO, USA
bianca.trinkenreich@colostate.edu

Abstract

Peer review is the primary mechanism through which software engineering (SE) research establishes quality and credibility, but it is also a social institution that shapes whose work and voices define the field. Yet many researchers increasingly experience this process as stressful, uneven, and misaligned with the goals of rigorous and meaningful scholarship. We analyzed open-ended responses from a community survey conducted ahead of ICSE 2026 FOSE to examine what SE researchers perceive as working well in peer review—and bringing joy—as well as what generates stress. Respondents value rigorous evaluation, intellectual stimulation, mentoring, and community support, but point to excessive submission volumes, time pressure, and inconsistent or superficial reviews as major sources of strain. Less experienced researchers and those with ten or fewer submissions in the past three years were about twice as likely as more experienced and higher-volume submitters to report that peer review is not working well. These patterns suggest that current practices disproportionately burden early career and most vulnerable researchers, highlighting the need to rebalance workload, accountability, and incentives to sustain a healthy SE research community.

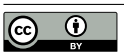
CCS Concepts

• **Software and its engineering**;

Keywords

peer review, software engineering research, reviewer workload

ACM Reference Format:
Júlia Condé Araújo, Marina Condé Araújo, Fabio Marcos de Abreu Santos, and Bianca Trinkenreich. 2026. The beauty and the pain of peer review in Software Engineering. In *2026 IEEE/ACM 48th International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3793657.3793872



ICSE-FoSE, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2479-4/2026/04
https://doi.org/10.1145/3793657.3793872

1 Introduction

The software engineering (SE) research community faces growing pressure. Longstanding concerns around "publish or perish" have intensified, reinforcing incentives that prioritize volume and novelty over lasting impact [6]. Increasing submission rates place sustained strain on the peer-review process, while prevailing evaluation norms continue to raise concerns about research relevance and researcher stress [1]. These challenges now intersect with the rapid adoption of generative AI, which may reshape SE research practices while introducing new ethical and governance risks [7]. Together, these dynamics raise questions about the sustainability and values of SE research.

In response, members of the SE community have called for reflection on research incentives and evaluation practices. In this paper, we contribute to this discussion by analyzing responses from a community survey conducted in late 2025 [5]. Our analysis examines aspects of peer review perceived as working well ("the beauty"), not working well ("the pain"), and the changes researchers consider necessary to improve the process ("healing the pain"), providing a basis for discussing how peer review norms might better support rigor, relevance, and researcher well-being.

The following research questions guided our study:

- RQ1 The beauty: What aspects of peer review are working well and bringing joy in SE research?
- RQ2 The pain: What aspects of peer review are not working well and bringing stress in SE research?
- RQ3 Healing the pain: What changes do researchers report as necessary to improve peer review in SE research?

In the remainder of this paper, we present our research design (Sec. 2), followed by the results (Sec. 3). After discussing threats to validity (Sec. 4), we conclude this paper with a discussion of the results (Sec. 5).

2 Research Design

In this section we lay out our research design as show in the Figure 1.

2.1 Data Collection and Preparation

The data analyzed in this paper originate from a community-facing survey designed by the ICSE 2026 FOSE organizers to elicit reflections on the state, practices, and future directions of SE research [5]. The questionnaire comprised 12 optional, open-ended questions, allowing participants to engage selectively with topics relevant to their experiences.

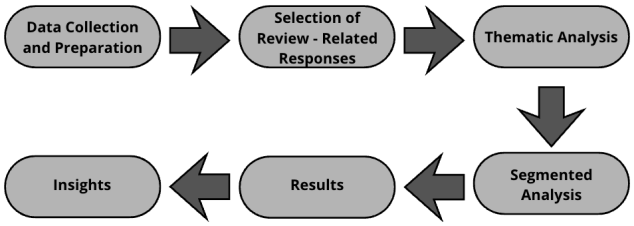


Figure 1: Fluxogram of the research design

The instrument included questions about what works well and not well in the SE research community, sources of joy and stress, and prioritized changes. In this paper, we focus on the subset of anonymized responses that referenced the term "review" to address our research questions: (RQ1) which aspects of peer review are perceived as working well and bringing joy, (RQ2) which aspects are perceived as not working well and experienced as stress, and (RQ3) what changes participants consider necessary to improve peer review in SE research.

Participation was anonymous, and participants were informed that their responses would be publicly shared via the FOSE Discord channel to support transparency and collective reflection. The survey was broadly disseminated, resulting in a convenience sample spanning different career stages, roles, and research backgrounds.

Our goal is not statistical generalization, but to surface recurring themes, tensions, and aspirations related to peer review in SE research, in order to inform discussion at FOSE.

2.2 Data Analysis

The authors collaboratively conducted qualitative analysis of the open-ended responses using an inductive, data-driven approach inspired by open coding [3]. Rather than applying a predefined coding scheme, the analysis focused on closely reading the responses and iteratively identifying recurring issues, perceptions, experiences, and suggested changes related to peer review in SE research.

We focused our analysis on open-ended survey responses that explicitly referenced the term "review" . In total, 279 participants completed the survey. Across the relevant open-ended questions, the number of non-blank analyzed responses was 30 (what works well), 4 (sources of joy), 87 (what does not work well), 57 (sources of stress), and 56 (suggested changes).

Emerging themes were progressively refined through comparison across responses, with attention to internal consistency and variation in how participants articulated similar positive and negative experiences. Throughout this process, the emphasis was on surfacing patterns grounded in participants' own language rather than producing an exhaustive or mutually exclusive code system.

The developing interpretations were then discussed among the authors. Through iterative discussions, we examined alternative readings of the data, clarified theme boundaries, and resolved disagreements through negotiated agreement.

To analyze how the perceptions of what is (and is not) working well in peer review differ according to individual characteristics, we segmented our sample based on *seniority* (less experienced: $\leq$ 5 years of experience vs. more experienced: $\geq$ 6 years of experience), *industry* (if the researcher's role is practitioner or from industry vs. academia (if the researcher's role is faculty, postdocs or student), and *peer-review exposure* (high: $\geq$ 11 paper submissions as (co-)author over the past three years vs. low $\leq$ 10 submissions over the past three years). Experience was dichotomized into less experienced (<= 5 years) and more experienced (>= 6 years) groups.

We applied the modified Haldane-Anscombe" (mHA) correction adding 0.5 to each term [8] to make possible to compute the odds ratio for Academic vs. Industry due to the low number of responses from industry.

Next, we calculated the odds ratio for the number of respondents who reported that peer review is (not) working well and the demographic information. We interpreted the results as follows:

- if **Odds Ratio = 1**, both groups are equally distributed for reporting that peer review is (or is not) working well.
- if **Odds Ratio > 1**, the likelihood for reporting that peer review is (or is not) working well is higher for the first group (in our case: more experienced, industry, high exposure to peer review in last three years).
- if **Odds Ratio < 1**, the likelihood for reporting that peer review is (or is not) working well is higher for the second group (in our case: less experienced, not from industry, low exposure to peer review in last three years).

3 Results

We analyzed survey responses related to the peer-review system using a two-level thematic coding approach. Responses were classified as describing peer review as working well or not working well and then organized into sub-categories capturing more specific experiences. Representative excerpts are presented throughout this section, with numbers in parentheses referring to anonymized Response IDs from the ICSE 2026 FOSE pre-survey raw data.

3.1 The beauty about peer review in SE (RQ1)

Our analysis of RQ1 identified two categories of aspects perceived as working well in peer review. In addition to describing effective mechanisms, participants explained how these aspects contribute to positive and rewarding experiences. We describe each category below using representative quotes.

3.1.1 Review quality & rigor. Participants mentioned that SE peer review is both rigorous and with strong attention to reproducibility. One respondent described SE as *"a strong community characterized by rigorous peer review, robust experimentation, and generally high-quality research output"* (192101946), while another noted that *"there are bars and criteria for the reviewers; the review process is, in general, detailed and rigorous"* (195861298).

Several participants highlighted the community's growing methodological maturity, particularly in its emphasis on empirical validation and causal reasoning. As one participant observed, SE research has moved toward by having *"rebuttals, major revisions or going to a journal [...], and also better empirical standards thanks to [review standards] [4]"* (195869297). Others noted the importance given by the SE community to reproducibility, artifact availability, and replication packages as reinforcing review quality and trust in published results (195856398).

From reviewing to intellectual stimulation. Reviewing papers, serving on committees, and engaging with new work were seen as joyful opportunities to broaden own expertise and learn from others' approaches. One respondent emphasized the joy of *"learning new things … by participating in a committee, by reviewing papers, or by getting hands-on with research"* (192295956), while another highlighted enjoyment in *"reading new papers during review"* when they are of high quality (192349636).

3.1.2 Mentoring and community support. A closely related theme is the constructive and developmental nature of feedback. Participants described SE peer review as embedded in a broader culture that prioritizes support, learning, and fairness. Overall, SE reviews were characterized by some participants as having *"high-quality, actionable, and generally valid critiques"* (192475387), *"relatively detailed and constructive"* (195811076). This orientation in peer review was perceived as mentoring and collective community support, benefiting researchers both as authors and reviewers. One participant highlighted that artifact evaluation roles, doctoral symposium, mentoring workshops, and junior program committee initiatives as mechanisms that *"help nurture early-career researchers and maintain a high standard of rigor and impact"* (192143242). Others emphasized a shared responsibility and the *"commitment to support each other … and help others to improve their work"* (196022336). One participant described the SE community as open to new ideas and emerging research areas, characterizing it as *"friendly and constructive in its ways to improve processes such as reviewing"* (195816320). Practices such as double-blind review and cautious approaches to AI use in reviewing were cited as reinforcing trust and inclusiveness (195756655).

Peer review as stewardship and advocacy. Participants expressed their satisfaction in helping authors improve their papers and in advocating for work they believed had merit. One participant described particular enjoyment in *"championing for acceptance if [they] believe a paper has merit"* and surprising authors *"with good news and the community with a really cool paper"* (192475387). Others connected this experience to mentoring, noting that reviewing and committee service *"gives [them] a chance to contribute to improving the review process itself, pushing for constructive, timely feedback."* (196044669).

3.2 The pain about peer review in SE (RQ2)

Our analysis of RQ2 identified two categories of aspects perceived as not working well in peer review, and related stressors. We describe each below with representative quotes.

3.2.1 Avalanche of paper submissions. The most prevalent concern related to the current publication and peer-review system, which participants widely described as overloaded and increasingly misaligned with community goals. Many participants explicitly linked ranking systems, venue prestige, and productivity metrics to a *"paper machine"* culture (192288951), arguing there are *"too much and too many low-quality papers… [which] is unsustainable"* (192161829). This environment was described as demoralizing and anxiety-inducing, reinforcing the perception that success depends more on strategic output than on meaningful contribution.

Participants repeatedly emphasized the sheer volume of submissions, noting that there are *"too many papers submitted and resubmitted"* (192082119). Others highlighted downstream consequences, observing that the reviewing load has become excessive and *"lessens the quality of the peer review process"* (192161829).

From Publication Volume to Time Pressure. Participants frequently described discomfort with the community's emphasis on publication volume over research quality, framing this as an ethical tension. One respondent criticized the pervasive *"numbers game,"* noting that *"it's just ridiculous to see how many papers some people publish per year"* (195853386).

This publication pressure was commonly experienced as temporal overload, particularly in reviewing and service work. As one participant noted, *"the ever growing amount of reviews (papers, grant proposals, etc.)"* is overwhelming (192867648). Others noted that these norms also extend to mentoring, describing the stress of advising students to prioritize flagship venues *"even if I think they should build a presence in their sub-community first"* (193204305).

3.2.2 Peer review uncertainty. Beyond volume, participants raised concerns about the quality and consistency of peer review. Reviews were often perceived as random, overly harsh, or superficial, with reviewers focusing on minor flaws rather than contribution or insight. One respondent observed that *"reviewers latch on to minor imperfections to reject more creative or novel papers"* (192134556), a pattern that discourages innovation and negatively affects student engagement in academia.

Concerns about bias and gatekeeping were also prominent. Several participants pointed to continuity and role concentration in reviewing and governance structures, noting that *"the same committee members appear repeatedly"* (192143242) and that committees may *"invite their friends to join"* (192109802), which some described as *"inbreeding within the community"* (192587847). In contrast, other participants attributed this continuity to structural constraints rather than intentional exclusion, emphasizing that conference service is *"voluntary hard work"* (192161434) and that it is *"often difficult to find qualified reviewers"* willing and able to serve at scale (192860086, 195869297).

From Uncertainty to Anxiety. Concerns about peer review quality were strongly associated with stress arising from evaluative interactions. Participants described frustration and exhaustion resulting from reviews perceived as inconsistent, superficial, or adversarial. One respondent detailed the difficulty of responding to reviews they viewed as factually incorrect, noting that *"reviewers invariably make up incorrect nonsense… any attempt to push back just makes people huffy and even more negative"* (192109802). Others emphasized the unpredictability of outcomes, observing that *"the review process sometimes seems to be arbitrary and luck-based"* (195802584).

Repeated cycles of rejection and resubmission further intensified this stress. For example, one participant described *"resubmitting papers with no guarantee that the reviewers will value upgrades made with respect to the prior version"* (192134556), while another reported sitting on *"an article that has been rejected five times, where only one review offered a proper reason"* (196078058).

Table 1: Odds ratio of reporting aspects that are (not) working well in peer review per personal characteristic

	Less experienced (<=5) vs. More experienced (>=6)	Academia vs. Industry	High peer-review exposure vs. Low peer-review exposure
Peer review is (not) working well (any reason)			
Peer review is working well	0.71	1.17	0.94
Peer review is not working well	2.26**	0.81	0.56**

Significance codes: * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Note: Odds ratio > 1 means that the first segment has greater chances of reporting the challenge than the second. Ratio < 1 means the opposite. The challenges were coded from the open question.

3.3 Segmented analysis

In addition to the thematic categorization described above, we analyzed the distribution of responses to examine how often respondents reported that peer review is—and is not—working well across key demographic groups, including career stage (less vs. more experienced), professional role (industry vs. academia), and recent exposure to peer review (low vs. high submission volume in the past three years), as defined in Sec. 2.2. Counts and proportions are reported as supplementary evidence to contextualize and triangulate the qualitative findings [2].

For reports that peer review is working well, no significant differences were observed, indicating similar positive assessments among less and more experienced respondents. In contrast, clear differences emerged for reports that peer review is not working well. Less experienced respondents had 2.21 times higher odds of reporting negative experiences than more experienced respondents, and those with lower recent submission volume (ten or fewer papers in the past three years) were about 1.8 times more likely to report negative assessments than those with higher submission exposure.

3.4 How to heal the pain? What needs to be changed in peer review (RQ3)

Across responses, participants claimed for rebalancing submission volume and incentives, reforming review processes and accountability, and redefining the role of conferences and journals.

3.4.1 Rebalancing submission volume and reviewer workload. A dominant theme was the need to reduce excessive submission volumes and better align reviewing efforts with authorship. Participants repeatedly emphasized that current expectations are unsustainable, leading to declining review quality and reviewer burnout. Several participants proposed enforcing stronger reciprocity between submitting and reviewing, such as requiring authors to review a fixed number of papers per submission. One participant argued that *"whoever submits to a journal or conference will need to engage in as many reviews as papers that are submitted"* (192732325), while another noted that reviewing far more papers than one submits is *"too much"* and fundamentally unfair (192349636). Others proposed structural mechanisms to slow down the publication cycle and discourage quantity-driven behavior. Suggestions included limiting the number of submissions per author (196061190) and shifting incentives in hiring and promotion to focus on a small number of high-quality contributions rather than long publication lists (193204305).

3.4.2 Improving accountability, transparency, and review quality. Participants also called for changes to increase accountability and consistency in reviews through greater transparency through open review models. Several participants advocated publishing reviews and reviewer identities, arguing that openness could discourage low-effort or arbitrary feedback. As one participant stated, *"open review, publish the name of reviewers"* (192288951), while another expressed hope that public reviews would lead reviewers to *"do a better job"* (195849456).

Beyond openness, participants emphasized the need for better reviewer training, oversight, and quality control. Some suggested that review quality should be monitored and that reviewers who consistently provide low-quality reviews should be excluded from committees (195860086). Others highlighted the need for clearer evaluation standards and stronger artifact requirements, arguing that *"no paper [should be accepted] without artifact evaluation"* to improve reproducibility and trust (195853386).

3.4.3 Rethinking the role of conferences and journals. Finally, many participants proposed rethinking the relationship between conferences and journals in SE research. A substantial subset argued that conferences *"should not mimic or replace journals"* (192250097), but instead focus on early feedback, discussion, and community building. Others proposed shifting most archival publications to journals and using conferences primarily for shorter presentations and discussion-oriented formats (192065107, 196022336).

Related suggestions included shorter papers with optional appendices (192134556), increased use of poster-style or interactive sessions (195838608), and clearer differentiation in review criteria between conferences and journals (196196304). These proposals were often motivated by a desire to reduce review burden, improve discussion quality, and restore conferences as spaces for intellectual exchange rather than high-pressure publication bottlenecks.

4 Threats to Validity

This study reflects a partial and situated view of how peer review is experienced in SE research. The data capture self-reported reflections from researchers who chose to respond to a community-facing FOSE survey. They emphasize articulated concerns and perceived changes rather than directly observed practices or outcomes.

The sample is subject to self-selection and survivorship bias. Researchers who are disengaged, overwhelmed, or have already withdrawn from the broader community are likely underrepresented. As a result, some forms of frustration, exclusion, or disengagement may be muted rather than amplified in our findings.

Our qualitative analysis is interpretive and exploratory by design. While themes were iteratively discussed among the authors, other readings of the data are possible. Finally, consistent with a position paper, the goal is not generalization or causal inference, but to surface shared tensions and blind spots that can inform collective reflection and discussion within the SE community.

5 Discussion and Conclusion

The results presented in this paper indicate that peer review in SE continues to be perceived as a central valued mechanism for ensuring research quality, while simultaneously being associated with increasing levels of strain. Rather than reflecting a contradiction, these findings suggest that peer review operates under conditions that challenge its original assumptions regarding scale, workload, and available time for careful evaluation.

Participants consistently describe positive aspects of peer review in terms of rigor, learning, and constructive engagement. Reviewing and receiving reviews are framed as opportunities for intellectual development and refinement of research contributions. The emphasis on empirical standards, reproducibility, and actionable feedback suggests shared expectations about sound SE research, indicating that peer review continues to fulfill its intended role in quality assurance and scholarly exchange.

At the same time, respondents report that the current volume of submissions places considerable pressure on both authors and reviewers. The cumulative effect of repeated submission cycles, compressed deadlines, and increasing service demands reduces the capacity for in-depth evaluation and limits opportunities for developmental feedback. Our segmented analysis shows that negative perceptions of peer review are more frequent among less experienced researchers and among those with lower recent submission exposure. This pattern suggests that the burdens associated with peer review are not evenly distributed and that researchers with less experience or fewer resources may be more exposed to its adverse effects.

Concerns related to uncertainty further contribute to these experiences. Participants report challenges arising from variability in review quality, inconsistencies in evaluation criteria, and limited transparency in decision-making processes. Under such conditions, peer review may be perceived as unpredictable, undermining confidence in the system even when reviewers act in good faith. Repeated rejections and resubmissions, particularly when feedback is perceived as insufficiently grounded or inconsistent across venues, amplify these concerns and contribute to cumulative stress.

Taken together, these observations suggest that current challenges in peer review are primarily structural rather than individual. The system relies on voluntary labor and shared norms within an incentive environment that encourages high submission rates.

The changes proposed by respondents reinforce this interpretation. Suggestions such as rebalancing, reviewing obligations, improving accountability mechanisms, and clarifying the respective roles of conferences and journals reflect an interest in preserving review quality rather than reducing selectivity or rigor. These proposals point toward the need for deliberate reflection on how peer review responsibilities are allocated and how evaluation processes can remain consistent and fair under increasing workload.

Building on these suggestions, we identify several directions for improving peer review that preserve human judgment while addressing structural strain:

- **Workload rebalancing and recognition:** More explicit alignment between submission volume and reviewing obligations, greater institutional recognition of reviewing and organizing labor, potentially including conference registration discounts as a way to acknowledge reviewing contributions.
- **Accountability and calibration:** Clearer expectations for review quality, lightweight mechanisms for feedback on reviews, and improved calibration across reviewers and venues to reduce variability in evaluation standards.
- **Process transparency and predictability:** Greater clarity in decision rationales and reduced uncertainty across repeated submission cycles.
- **Supportive use of Generative AI in peer review:** Participants' concerns about workload and inconsistency suggest opportunities for GenAI to assist (rather than replace) human reviewers and program chairs. Potential roles include checking artifacts, summarizing reviewer discussions and meta-reviews, flagging missing justifications, and unconstructive tone. Importantly, such tools should be designed to augment human judgment, preserve reviewer agency, and avoid automating evaluative decisions themselves [7].

In conclusion, this study highlights that peer review in SE remains a valued component of the research process, yet facing growing pains related to scale and sustainability. Addressing these challenges requires recognizing peer review as a collective practice shaped by institutional incentives and community norms. By grounding future discussions and potential adjustments in the experiences reported by researchers, the community can work toward maintaining review processes that support both research quality and long-term participation in SE research.

References

[1] Vahid Garousi, Markus Borg, and Markku Oivo. 2020. Practical relevance of software engineering research: synthesizing the community's voice. *Empirical Software Engineering* 25, 3 (2020), 1687–1754.

[2] David R Hannah and Brenda A Lautsch. 2011. Counting in qualitative research: Why to conduct it, when to avoid it, and when to closet it. *Journal of Management Inquiry* 20, 1 (2011), 14–22.

[3] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

[4] Paul Ralph, Nauman bin Ali, Sebastian Baltes, Domenico Bianculli, Jessica Diaz, Yvonne Dittrich, Neil Ernst, Michael Felderer, Robert Feldt, Antonio Filieri, et al. 2020. Empirical standards for software engineering research. *arXiv preprint arXiv:2010.03525* (2020).

[5] Margaret Storey and Andre van der Hoek. 2025. Community Survey for ICSE 2026 Future of Software Engineering: Toward a Healthy Software Engineering Community. doi:10.5281/zenodo.18217798 Zenodo repository record.

[6] Margaret Anne Storey. 2019. Publish or perish: Questioning the impact of our research on the software developer. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE, 2–2.

[7] Bianca Trinkenreich, Fabio Calefato, Geir Hanssen, Kelly Blincoe, Marcos Kalinowski, Mauro Pezzè, Paolo Tell, and Margaret-Anne Storey. 2025. Get on the train or be left on the station: Using llms for software engineering research. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*. 1503–1507.

[8] Frank Weber, Guido Knapp, Katja Ickstadt, Günther Kundt, and Änne Glass. 2020. Zero-cell corrections in random-effects meta-analyses. *Research Synthesis Methods* 11, 6 (2020), 913–919.