A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences

Ivan Stelmakh[1], **Nihar B. Shah**[1], **Aarti Singh**[1], **Hal Daumé III**[2,3]
[1]School of Computer Science, Carnegie Mellon University
[2]University of Maryland, College Park
[3]Microsoft Research, New York
{stiv,nihars,aarti}@cs.cmu.edu, me@hal3.name

Abstract

Conference peer review constitutes a human-computation process whose importance cannot be overstated: not only it identifies the best submissions for acceptance, but, ultimately, it impacts the future of the whole research area by promoting some ideas and restraining others. A surge in the number of submissions received by leading AI conferences has challenged the sustainability of the review process by increasing the burden on the pool of qualified reviewers which is growing at a much slower rate. In this work, we consider the problem of reviewer recruiting with a focus on the scarcity of qualified reviewers in large conferences. Specifically, we design a procedure for (i) recruiting reviewers from the population not typically covered by major conferences and (ii) guiding them through the reviewing pipeline. In conjunction with the ICML 2020 — a large, top-tier machine learning conference — we recruit a small set of reviewers through our procedure and compare their performance with the general population of ICML reviewers. Our experiment reveals that a combination of the recruiting and guiding mechanisms allows for a principled enhancement of the reviewer pool and results in reviews of superior quality compared to the conventional pool of reviews as evaluated by senior members of the program committee (meta-reviewers).

1 Introduction

Over the last few years, Machine Learning (ML) and Artificial Intelligence (AI) conferences have been experiencing rapid growth in the number of submissions: for example, the number of submissions to AAAI and NeurIPS — popular AI and ML conferences — more than quadrupled in the last five years. The explosion in the number of submissions has challenged the sustainability of the peer-review process as the number of *qualified* reviewers is growing at a much slower rate (Sculley, Snoek, and Wiltschko 2019; Shah 2019b). While especially prominent in ML and AI, the problem is present in many other fields where "*submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint*" (McCook 2006).

The disparity between growth rates of the submission and reviewer pools increases the burden on the reviewers, thereby putting a severe strain on the review process. According to the president of the International Conference on Machine Learning (ICML) board John Langford (Langford 2018), *"There is significant evidence that the process of reviewing papers in machine learning is creaking under several years of exponentiating growth."* Hence, it is important to increase the number of qualified reviewers in the system to keep up with the growing number of submissions.

When the size of a conference is small, program chairs can extend the pool of reviewers by manually selecting new reviewers among researchers who have enough expertise in the area. The selection can be guided by the program chairs' understanding of who might be a good reviewer or by personal recommendations made by other senior members of the program committee. In what follows, we refer to the pool of reviewers manually constructed by the program chairs as the curated pool. However, with a massive increase in the scale of the conference, such a manual addition to the curated pool does not allow bringing in enough reviewers to cover the demand of the conference. The program chairs must then rely on alternative ways of reviewer recruiting.

With this motivation in mind, in the present paper we aim to design and evaluate modifications to the reviewer recruiting process that simultaneously address two challenges:

- **Challenge 1.** To avoid overloading reviewers, conferences need to find new sources of reviewers as there are not enough curated reviewers to review all papers.
- **Challenge 2.** Conferences need to ensure that newly added reviewers do not compromise the quality of the process, that is, are able to write reviews of quality at least comparable to the curated reviewer pool.

In the past, conference organizers have been trying to expand the reviewer pool by relaxing the qualification bar, that is, by allowing researchers who meet some minimal requirements such as having one or two relevant publications to join the pool of reviewers without further screening. For example, 1176 out of 3242 (that is, 36%) of the reviewers in the NeurIPS 2016 conference were recruited by requesting authors of each submission to name at least one author who is willing to become a reviewer, and 70% of these reviewers were PhD students: researchers at very early stages of their careers (Shah et al. 2018). Such practices have now become conventional and are adopted by many other conferences, including a flagship conference in artificial intelligence AAAI that in 2020 invited self-nominated individuals with publication history in top venues, and in 2021 requires authors of submissions to be willing to become reviewers on request.

While the aforementioned innovations allow to enlarge the reviewer pool, little scientific evidence exists on the quality of reviews written by reviewers recruited through these novel procedures. NeurIPS 2016 compared the reviews written by curated reviewers with reviews sourced from authors of submissions in terms of numeric scores (overall score and several criteria scores) and inter-reviewer agreement (Shah et al. 2018). The analysis did not reveal a significant difference between populations, only showing that author-sourced reviews were slightly harsher in scoring the clarity of submissions. However, we note that this analysis only operates with scores given by reviewers and does not address the *quality* of reviews — perhaps the most important metric for success of the conference peer-review process — which is largely determined by the textual part of the review. Other works provide anecdotal and empirical evidence that junior reviewers are more critical than their senior counterparts (Mogul 2013; Toor 2009; Tomiyama 2007) and that "*graduate students seem to be unable to provide very useful comments*" (Patat et al. 2019). Thus, while the methods employed by leading conferences address the first challenge, it remains unclear if and how they address the second challenge of high quality reviewing.

In this work, in conjunction with the review process of ICML 2020 we conduct a threefold experiment:

- First, we recruit reviewers from the population not typically covered by the reviewer-selection process of major conferences. In that, we target the population of very junior researchers with limited or no publication/reviewing history most of whom do not pass the recruiting filters of ICML. Conceptually, in contrast to the standard approach of selecting reviewers based on some proxy towards reviewing ability (e.g., prior publication and reviewing history), we evaluate candidates' abilities to review in an auxiliary peer-review process organized for the experiment.
- Second, we add a select set of reviewers recruited through our experiment to the reviewer pool of the ICML conference and guide them through the peer-review process by offering mentoring.
- Finally, we evaluate the performance of these novice reviewers by comparing them with the general population of the ICML reviewer pool on multiple aspects. In doing so, we augment the past analysis of Shah et al. (2018) by using an explicit measure of review quality (evaluated by meta-reviewers) in addition to indirect proxies.

An important aspect of our experiment is that most of the reviewers brought to the reviewer pool through our experiment would not have been considered in standard ways of recruiting. Hence, our experiment offers a principled way to enlarge the reviewer pool. As a by-product, the new pool of reviewers contributes to diversity of peer review resonating with the virtues such as increased scrutiny and variety of opinions outlined by Garisto (2019). Moreover, we offer the new reviewers a more guided introduction to the reviewing process which is known to help novice reviewers to write better reviews (Patat et al. 2019) and improve their own writing skills (Kerzendorf et al. 2020). From the perspective of training reviewers, our experiment is conceptually similar to the initiative of *Journal of Neuroscience* (Picciotto 2018) and SIGCOMM conference (Feldmann 2005) that attempt to help novices in becoming reviewers.

This work also falls in the line of empirical works that study various behavioral aspects of human computation, including the impact of competitive (Levy and Sarne 2018) and impartial (Kotturi et al. 2020) framing of the task on performance of the human agents. Additionally, it continues another direction of research (Kurokawa et al. 2015; Xu et al. 2019; Lian et al. 2018; Stelmakh, Shah, and Singh 2019) that aims at improving the conference peer review.

The rest of the paper is structured as follows. In Section 2 we discuss the methodology of each component of the novice-reviewer experiment. We then present the main results in Section 3. Finally, in Section 4 we conclude the paper with a discussion of various aspects of the experiment.

2 Methodology

In this section we discuss the setup of our experiment. We introduce the selection and mentoring mechanisms and explain the methodology of evaluation of reviewers recruited through our experiment in the ICML 2020 conference.

2.1 Selection Mechanism

The high-level idea of our selection mechanism is to pretest abilities of candidates to write high-quality reviews. To this end, we frame the experiment as an auxiliary peer-review process that mimics the pipeline of the real ML conferences as explained below and ask participants to serve as reviewers for this conference. Let us now describe the experiment in detail by discussing the pools of participants and papers, the organization of the auxiliary review process, and the selection criteria we used to identify the best reviews whose authors were invited to join the ICML reviewer pool.

Papers We solicited 19 anonymized preprints in various sub-areas of ML from colleagues at various research labs, ensuring that authors of these papers do not participate in the experiment as subjects. Some ML and AI conferences publicly release reviews for accepted/submitted papers, making these papers inappropriate for our experiment as our goal is to elicit independent reviews from participants. Thus, we used only those papers that did not have reviews publicly available. The final pool of papers consisted of working papers, papers under review at other conferences, workshop publications and unpublished manuscripts. The papers were 6–12 pages long excluding references and appendices (a standard range for many ML conferences) and were formatted in various popular journals' and conferences' templates with all explicit venue identifiers removed.

Participants Since we had a small quota, in this positional experiment we limited the target study population to graduate students or recent graduates of five large, top US universities (CMU, MIT, UMD, UC Berkeley and Stanford). To recruit participants, we messaged mailing lists of these universities and targeted master's and junior PhD students working in ML-related fields, the invitation also propagated to a small number of students outside of these schools through the word of mouth. The recruiting materials contained an invitation to participate in the ICML reviewer-selection experiment. Specifically, we notified participants that they will need to review one paper and that those who write strong reviews will be invited to join the the ICML reviewer pool. Being a reviewer in the top ML conference is a recognition of one's expertise and we envisaged that this potential benefit is a good motivation for junior researchers to join our experiment. As a result, we received responses from 200 candidates, more than 90% of whom were students/recent graduates from the aforementioned schools. All of these candidates were added to the pool of participants without further screening. We provide additional discussion of the demography of participants in Section 4.

Auxiliary peer-review process The selection procedure closely followed the initial stages of the standard double-blind ML conference peer-review pipeline. First, we asked participants to indicate their preferences in what papers they would like to review by entering bids that take the following values: "Not Willing", "In a Pinch", 'Willing' and "Eager". Thirteen participants did not enter any bids and were removed from the pool. Based on these bids, we assigned one paper each to all remaining participants, where we tried to satisfy reviewer bids, subject to a constraint that each paper is assigned to at least 8 reviewers. As a result, 186 participants were assigned to a paper they bid either "Willing" or "Eager" and 1 participant was assigned to a paper they bid "In a Pinch" (this participant did not bid "Eager" or "Willing" on any paper).

Finally, we instructed participants that they should review the paper as if it was submitted to the real ICML conference with the exception that the relevance to ICML, formatting issues (e.g., page limit, margins) and anonymity issues should not be considered as criteria. To help participants in writing their reviews, we provided reviewer guidelines (included in supplementary materials on the first author's website) that discuss the best practices of reviewing. We gave participants 15 days to complete the review and then extended the deadline for 16 more days to accommodate late reviews as our original deadline interfered with the final exams at various US universities and the US holiday period.

Selection of participants Out of 187 participants who were assigned a paper for review, 134 handed in the reviews (response rate of 71.7%). Upon receipt of reviews, we removed numeric scores given by participants to the papers and relied on the combination of the following approaches to identify individuals to be invited to join the ICML reviewer pool.

- **Internal evaluation** We analyzed reviews for all papers falling in the study team members' areas of expertise.
- **External evaluation** We called upon an independent domain expert to help with papers that are outside of the study team members' areas of expertise.
- **Author evaluation** We asked authors of papers used in the experiment to rate/comment on the qualities of reviews. Authors of 14 of the 19 papers responded to our request.

Combining feedback coming from these sources, we eventually invited 52 participants whose reviews received excellent feedback from all evaluators who read the review to join the ICML reviewer pool and all of them accepted the invitation. For the rest of the paper, we will refer to these reviewers as EXPERIMENTAL reviewers.

2.2 Mentoring Mechanism

Throughout the conference review process, the EXPERIMENTAL reviewers were offered additional mentorship:

- The reviewers were provided with a senior researcher as a point of contact, and were offered to ask any questions on reviewing at any point of the process. There were several questions asked and answered as a part of the mentorship.
- The reviewers were provided with examples on various parts of the process, for instance, on how to lead a discussion among the reviewers.
- At the point when the initial reviews were submitted, certain issues were identified that were common across many reviews from the EXPERIMENTAL pool (e.g., many reviews were initially written about the authors rather than the paper). The EXPERIMENTAL reviewers were requested to address these issues.
- The EXPERIMENTAL reviewers were sent a few more reminders than the conventional reviewers.

The amount of time and effort in the mentorship was equal to about half the time and effort for a meta-reviewer's job.

2.3 Methodology of Evaluation

The main pool of the ICML 2020 reviewers was recruited through a combination of conventional approaches and consisted of 3,012 reviewers[1] belonging to two disjoint groups. The first group, which we refer to as CURATED, made up about 68% of the main pool and included reviewers who were invited by program chairs based on satisfaction of at least one of the following criteria: (i) several years of reviewing and publishing experience for top ML venues, (ii) above-average performance in reviewing for NeurIPS 2019 or (iii) personal recommendation by a meta-reviewer. The remaining 32% of reviewers constituted the second group that we call SELF-NOMINATED: this group comprised individuals who self-nominated and satisfied the selection criteria of (i) having at least two papers published in some top ML venues, and (ii) being a reviewer for at least one top ML conference in the past. On average, the CURATED group consisted of more senior researchers while the SELF-NOMINATED pool mostly comprised researchers at early stages of their careers.

In the sequel, we compare the performance of 52 EXPERIMENTAL reviewers who joined the ICML reviewer pool through our experiment with the performance of the reviewers from the main pool. Let us now discuss some important details of the evaluation.

Affiliation caveat 51 out of 52 EXPERIMENTAL reviewers recruited through our selection procedure are current master's and PhD students or recent graduates of the aforementioned universities (one reviewer is a graduate of another

[1] Some reviewers who initially accepted the invitation dropped out in the early stages of the review process and are not included in this number and in the subsequent analysis.

US school), whereas reviewers from the main pool represent universities as well as private companies, government organizations, non-profits and more, from all over the world. Hence, the reviewers in the main pool have different backgrounds from the EXPERIMENTAL reviewers and this difference can serve as an undesirable confounder (orthogonal to the selection procedure and mentoring) in our analysis.

To counteract this confounding factor, we identify a subset of the main pool of reviewers, whom we call COLLEAGUE reviewers. The COLLEAGUE group comprises 305 reviewers from the main pool who share an affiliation (i.e., email domain or affiliation listed on the conference management system) with the 5 schools mentioned above. In our evaluations subsequently, we additionally juxtapose the EXPERIMENTAL reviewers to this group to evaluate how they compare to reviewers of similar background, thereby alleviating the affiliation confounder.

Metrics and tools of comparison We use a set of indirect indicators of review quality (e.g., review length and discussion participation) as well as direct evaluations of review quality made by meta-reviewers — senior reviewers, each of whom is in charge of overseeing the review process for approximately 20 submissions — of the ICML conference. To quantify significance of the difference in these metrics, we use the permutation test (Fisher 1935), treating each paper-reviewer pair as a unit of analysis. Error bars presented in figures below represent bootstrapped 90% confidence intervals unless stated otherwise.

Finally, throughout the conference meta-reviewers were calling upon additional external reviewers to help with some submissions or asking reviewers from the main pool to review additional papers; these paper-reviewer pairs are not included into comparison because new reviewers typically had less time to complete reviews.

3 Evaluation

In the previous section we described our approach towards recruiting novice reviewers and mentoring them. In this section we move to the real ICML conference and evaluate the benefit of the proposal by juxtaposing the performance of EXPERIMENTAL reviewers to the main reviewer pool which consists of SELF-NOMINATED and CURATED reviewers, some of whom belong to the group of COLLEAGUE reviewers. For this, we compare performance of reviewers at different stages of the review process: bidding, reviewing (in-time submission, review length, self-assessed confidence and others) and discussion (activity, attention to the author feedback). Finally, we complement the comparison by overall evaluation of the review quality made by meta-reviewers.

Table 1 summarizes the results of comparison of EXPERIMENTAL reviewers with reviewers from the main pool; subsequently, we will present a more detailed analysis with breakdown by reviewer groups. The main message of Table 1 is that from various angles the reviews written by EXPERIMENTAL reviewers are comparable to or sometimes even better than reviews written by reviewers from the main pool. With this general observation, we now provide details and background for select rows of Table 1. A more detailed analysis and statistics are given in the full technical report (Stelmakh et al. 2020).

Bidding activity (Row 1 of Table 1) Algorithms for automated paper-reviewer matching significantly rely on reviewer bids. Hence, activity of reviewers in the bidding stage is crucial to ensure that submissions are assigned to reviewers with appropriate expertise. To give matching algorithms enough flexibility, ICML program chairs requested reviewers to positively bid (i.e., indicate papers they are “Willing” or “Eager” to review) on at least 30-40 submissions (out of approximately 5,000 submitted for review).

Figure 1 compares mean numbers of positive and non-negative (positive bids and “In a Pinch”) bids made by reviewers from different groups. We observe that EXPERIMENTAL reviewers are more active than other categories of reviewers with qualification that the difference with SELF-NOMINATED reviewers ($\Delta_{\text{positive}} = 3.4,\ \Delta_{\text{non-negative}} = 5.0$) is not statistically significant at the 0.05 significance level.[2]

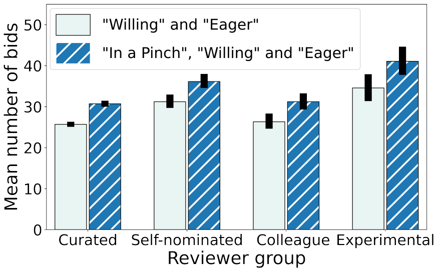

Figure 1: Mean number of positive/non-negative bids per reviewer. EXPERIMENTAL reviewers positively bid on more papers than reviewers from each of the comparison groups.

Timely review submission (Row 2 of Table 1) A typical conference timeline is very tight and it is crucial that reviewers complete their reviews in a timely manner. Figure 2 juxtaposes engagement ratios — fractions of reviewers who submitted at least one review by a given date — of different reviewer groups. Again, we observe the trend of junior reviewers being more active than their senior counterparts with EXPERIMENTAL reviewers achieving the highest engagement ratio. Note that review submission deadline of the ICML conference was extended twice. The initial deadline on day X was extended well in advance and hence the completion rate on day X is very low, so in Table 1 we use data for the deadline on day X+3.

Finally, we need to qualify that in this section we compare the fraction of engaged reviewers instead of perhaps a more natural choice of completion rate (the number of submitted reviews divided by the number of assigned papers). The rationale behind our choice is the difference in the reviewer load between categories. As we discuss in Section 4, EXPERIMENTAL reviewers had a reduced load compared to other reviewers and this difference makes the completion rate artificially favourable to EXPERIMENTAL reviewers (we compare completion rates in the full technical report).

[2] We also observe that SELF-NOMINATED reviewers are more active than CURATED reviewers.

Criteria (R = reviewer, P = paper)	Range	Experimental	Main Pool	p-value
Mean number of positive bids per R*	$[0, 5052]$	34.6	27.4	.043
Frac. of Rs with > 0 completed reviews by the deadline*	[0, 1]	0.92	0.81	.041
Mean review length (in symbols)*	$[0, \infty)$	4759	2858	$< .001$
Mean initial overall score given by Rs	$[1, 6]$	3.34	3.25	.373
Mean self-reported confidence	$[1, 4]$	3.05	3.03	.841
Mean self-reported experience*	$[1, 4]$	2.83	2.98	.026
Frac. of (P, R) pairs with R active in P discussion*	$[0, 1]$	0.68	0.58	.033
Frac. of (P, R) pairs with review updated after rebuttal*	$[0, 1]$	0.61	0.43	$< .001$
Mean review quality evaluated by meta-R*	$[1, 3]$	2.26	2.08	$< .001$

Table 1: Performance comparison of reviewers from the main pool and EXPERIMENTAL reviewers on various criteria. Asterisks indicate criteria with significant difference at the level 0.05.

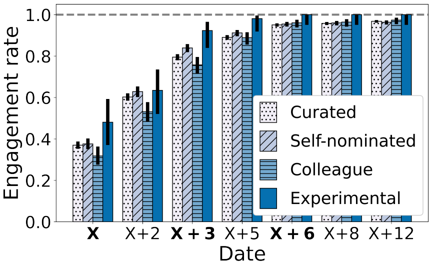

Figure 2: Fraction of engaged reviewers. Bold labels indicate dates at which deadlines were set with the original deadline on day X and two extensions. 90% confidence intervals are computed using the Wilson method (Wilson 1927). EXPERIMENTAL reviewers are consistently more engaged than reviewers from each of the comparison groups.

Review length (Row 3 of Table 1) We continue the analysis by juxtaposing the lengths of textual comments submitted by reviewers in Figure 3. We observe that different categories of reviewers from the main pool appear to write reviews of comparable length whereas EXPERIMENTAL reviewers write considerably longer reviews. The distribution of lengths of reviews written by reviewers from the main pool is very similar to that of several major ML conferences (Beygelzimer et al. 2019), and thus EXPERIMENTAL reviewers produced longer reviews than standard in the field.

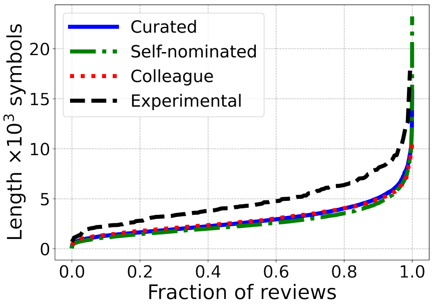

Figure 3: Distribution of review lengths. EXPERIMENTAL reviewers write longer reviews than other reviewers.

Rebuttals and discussion (Rows 7 and 8 of Table 1) The review process of ICML allows authors to respond to initial reviews written for their papers by submitting a short rebuttal that is followed by a private discussion between reviewers and the meta-reviewer. Past analysis (Shah et al. 2018; Gao et al. 2019; Gurevych, Miyao, and Cardie 2018) provide mixed evidence of the usefulness of rebuttals, and in this work we do not aim to judge the overall efficacy of the rebuttal process. However, in order for the rebuttal or discussion to change the reviewer's opinion, reviewers at the very least need to consider the rebuttal and be engaged in the discussion and we now investigate this aspect, conditioning on papers whose authors supplied a response to initial reviews.

Figure 4 compares the fractions of paper-reviewer pairs such that the reviewer posted at least one message in the discussion thread (discussion activity rate, left bars) / updated the textual review after the rebuttal (review update rate, right bars). We note that in both dimensions EXPERIMENTAL reviewers are more active than other categories of reviewers.[3]

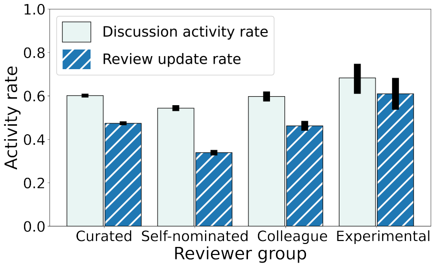

Figure 4: Activity in the last stage of the review process. EXPERIMENTAL reviewers participate in the discussion and update textual reviews more often than other reviewers.

Review quality (Row 9 of Table 1) So far we have observed that EXPERIMENTAL reviewers are more active in all stages

[3] Interestingly, Figure 4 shows that SELF-NOMINATED reviewers are less engaged in the last stage of the review process than senior CURATED reviewers. This observation suggests that the relative engagement of junior SELF-NOMINATED reviewers decreases as the review process progresses. We do not see this in the EXPERIMENTAL reviewers and hypothesize that more tailored mentoring leads to a consistent engagement of EXPERIMENTAL reviewers.

of the review process than main reviewer pool. However, the comparisons above do not decisively answer the question of *quality* of reviews written by the new reviewers. To bridge this gap, we now report the evaluations of review quality made by meta-reviewers. At the end of the review process, meta-reviewers were asked to evaluate the quality of each review on a 3-point Likert item with the following options: "Failed to meet expectations" (Score 1), "Met expectations" (Score 2), "Exceeded expectations" (Score 3). Importantly, meta-reviewers were not aware of the group affiliation of reviewers. For Table 1, we compare mean scores between different categories of reviewers.

Figure 5 visualizes the fraction of reviews below and above the expectations of meta-reviewers within each group. Observe that reviews written by EXPERIMENTAL reviewers exceed expectations of meta-reviewers more often than reviews of CURATED and SELF-NOMINATED reviewers (conditioning on the affiliation does not impact the comparison). Figure 5 also shows that EXPERIMENTAL reviewers produced substandard reviews less often than other reviewers.

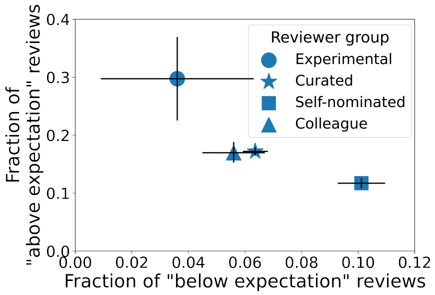

Figure 5: Evaluation of review quality by meta-reviewers. The closer the point to the upper-left corner, the better. EXPERIMENTAL reviewers dominate other groups of reviewers.

4 Discussion

In this work we designed and executed an experimental procedure for novice reviewer recruiting and guiding with a goal to address scarcity of qualified reviewers in large conferences. We evaluated the results of the experiment by juxtaposing the performance of the new reviewers to the traditional reviewer pool in the real ICML 2020 conference. We now provide additional discussion of the recruiting and evaluation procedures and suggest directions for future work. For more comprehensive discussion, please refer to the full technical report (Stelmakh et al. 2020).

4.1 Discussion of the Recruiting Experiment

We begin from some important aspects of the auxiliary peer-review process we used to recruit reviewers. First, we perform the analysis of the demography of participants. We then mention another potentially useful aspect of the experiment related to reviews written in the auxiliary review process.

Demography of participants Recall that self-nominated individuals had to pass a publication and reviewing filters (see Section 2.3 for details) to join the SELF-NOMINATED reviewer pool of ICML 2020. We now test whether the subjects of our experiment satisfy these criteria. Table 2 comprises relevant demographic information for 134 subjects of the experiment who completed the participation (that is, submitted the review of the assigned paper) and for 52 subjects who were eventually invited to join the ICML 2020 EXPERIMENTAL reviewer pool. Importantly, this demographic information was hidden from evaluators who performed the selection. Observe that most of the participants of our experiment (including those that were selected to join the ICML reviewer pool) do not pass at least one of the filters mandatory for the SELF-NOMINATED reviewers. Similarly, none of the participants was invited to join the CURATED reviewer pool. Therefore, we conclude that most of the participants of our experiment would not have been invited through conventional ways of reviewer recruitment.

	ALL	INVITED
TOTAL NUMBER	134	52
WITH PRIOR REVIEW EXPERIENCE	36%	37%
WITH PUBLICATIONS	72%	77%
PASS REVIEWING FILTER	21%	21%
PASS PUBLICATION FILTER	24%	23%
PASS SELF-NOMINATED FILTERS	**13%**	**12%**

Table 2: Demography of subjects of our experiment.

Reviews As a byproduct of our experiment, authors of manuscripts we used in the auxiliary peer-review process received a set of reviews. They generally admitted a high quality of reviews: several authors mentioned that reviewers found some errors/important typos in their papers or suggested some ways to improve the presentation, and an author of one paper says: "*Very high quality reviews, in my opinion. Most of them [...] are clearly more detailed and more useful than reviews we received at [another top ML venue]*". The positive feedback from authors hints that a large scale version of our experiment can give researchers a set of useful reviews before they submit a paper to a real conference, potentially decreasing the load on the actual conferences.

4.2 Discussion of the Evaluation

We now mention some aspects important for interpretation of the results of the comparison between EXPERIMENTAL reviewers and the main reviewer pool of ICML 2020.

Aspect 1. Assignment procedure Each paper submitted to the ICML conference was first automatically assigned to 3 reviewers from the main pool, two of whom were CURATED reviewers and one was a SELF-NOMINATED reviewer. In that, we tried to satisfy reviewer bids and optimize for the notion of textual similarity (Charlin and Zemel 2013) between submissions and assigned reviewers, subject to a requirement that each reviewer is assigned at most 6 papers (a small set of reviewers requested a lower quota) and under a fairness constraint (Stelmakh, Shah, and Singh 2018). After that, each EXPERIMENTAL reviewer was manually assigned

	MAIN POOL	EXPERIMENTAL
# REVIEWERS	3012	52
MEAN REVIEWER LOAD	5.4	3.0
FRAC. OF POSITIVE BIDS	0.88	0.99
MEAN SIMILARITY	0.77	0.88

Table 3: Assignment quality. "Fraction of positive bids" represents a fraction of paper-reviewer pairs in the assignment such that the reviewer has bid positively on the paper. Similarities between papers and reviewers take values in the interval $[0, 1]$. 706 reviewers in the main pool and 4 EXPERIMENTAL reviewers did not have similarities computed and are excluded from the computation of mean similarity.

to 3 submissions as a 4th reviewer. All assignments were finally adjusted by meta-reviewers before being released to the reviewers. Given the small number of EXPERIMENTAL reviewers and the large number of submissions, the constraints we had to satisfy in the manual assignment were mild as compared to the main assignment. Hence, EXPERIMENTAL reviewers could receive submissions that better fit their expertise than reviewers from the main pool.

Table 3 compares several metrics of assignment quality across reviewers from the main pool and EXPERIMENTAL reviewers. First, we note that EXPERIMENTAL reviewers were intentionally assigned less papers than reviewers from the main pool to ensure a gentle introduction to the review process. However, we underscore that in this study we aim to show that reviewers from the population not covered by current recruiting methods can usefully augment the reviewer pool given some special treatment (careful recruiting, reduced load and mentoring). Hence, we do not consider the difference in load to be a confounder in our comparisons.

Second, EXPERIMENTAL reviewers were assigned to papers they positively bid on and to papers with high textual similarity more often than reviewers from the main pool. Hence, we caveat that the quality of the assignment differs significantly between reviewers from the main pool and EXPERIMENTAL reviewers, introducing a confounding factor. On the other hand, the difference in the assignment quality may in part be due to the difference in bidding activity (see Section 3): a large number of positive bids made by EXPERIMENTAL reviewers gives more flexibility in satisfying them, thereby increasing the quality of the assignment. It will be of interest to investigate, in any future larger scale studies, whether a larger number of EXPERIMENTAL reviewers also continue to have such a higher quality of assignment due to higher bidding activity, and if not, then it will be of interest to observe how that impacts the other metrics.

Aspect 2. Quality evaluation Following a standard approach in AI/ML conferences that conduct a survey of meta-reviewers on the review quality, when asking meta-reviewers to evaluate the quality of reviews, we left it for the meta-reviewers to decide on their expectations and did not precisely define the term "quality". As a result, different meta-reviewers could have different standards in mind, leading to some inconsistency in evaluations. To account for this issue, in the full technical report we complement the above analysis of review quality and other metrics by restricting attention to submissions that had at least one EXPERIMENTAL reviewer assigned (by doing so we equalize the sets of meta-reviewers who rate EXPERIMENTAL reviewers and other categories of reviewers). Importantly, this analysis leads to the same conclusions as the analysis we described in Section 3.

Another related caveat is that the absence of a well-defined notion of quality could result in the substitution bias in meta-reviewers' judgments. For example, meta-reviewers' evaluations could be driven by the length of the review or some other computationally inexpensive, but sub-optimal, proxy, resulting in a biased evaluation of quality.

Aspect 3. The role of reviewers With the above caveats, the experiment demonstrated that EXPERIMENTAL reviewers are comparable to and sometimes even better than reviewers recruited in conventional ways in terms of various metrics analyzed in Section 3. However, we qualify that this observation absolutely does not imply that EXPERIMENTAL reviewers can entirely substitute the pool of experienced reviewers. Instead, we conclude that if recruited and mentored appropriately, EXPERIMENTAL reviewers can form a useful augmentation to the traditional pool. The EXPERIMENTAL and experienced reviewers may have different strengths that can be combined to achieve an overall improvement of the review quality. For instance (Shah 2019a), EXPERIMENTAL and, more generally, junior reviewers can be used to evaluate nuanced technical details of submissions while senior researchers can focus on the broader picture and more subjective criteria (e.g., impact) where their expertise is crucial.

4.3 Future Work

An important direction for future work is a design of a scalable version of the proposed selection procedure. The current selection pipeline requires an amount of work equivalent to 2 to 4 days of the conference workflow chair's work and 2 to 4 hours of the conference program chairs' work to execute the experiment. A more autonomous version of the procedure is needed if we wish to significantly increase the number of reviewers recruited through the proposed way.

Next, it would be interesting to compare EXPERIMENTAL reviewers with the main reviewer pool in a larger scale study that would enable a deeper analysis of textual reviews written by different reviewer groups.

Another important direction is a principled design of a mentoring protocol to support novice reviewers. Since ML/AI conferences have hundreds of meta-reviewers, it may be prudent to assign some meta-reviewers as mentors for junior reviewers and reduce their meta-reviewer workload accordingly. Future editions could also involve sharing more material on how to review with reviewers (e.g., Köhler et al. 2020) and holding webinars with Q&A sessions.

Finally, the feedback from the EXPERIMENTAL reviewers was that it was helpful for them to experience and gain insights into the review process, which will also help in their own research dissemination in the future. It would be interesting to measure the impact of the guided introduction to the review process in the early stages of career on the future trajectory of the individuals as researchers and reviewers.

Acknowledgments

We thank authors of papers used in the selection experiment for contributing their works to this study and for helping us with identifying strong reviews. We also thank Devendra Singh Chaplot who served as a domain expert in the external evaluation of reviews. We are grateful to the support team of the Microsoft Conference Management Toolkit (CMT) who hosted our selection experiment and helped with many customization requests during the experiment and the actual ICML 2020 conference. Finally, we appreciate the efforts of all reviewers and meta-reviewers involved in the ICML 2020 review process. This study was approved by Carnegie Mellon University Institutional Review Board.

This work was supported in part by NSF CAREER award 1942124 and in part by NSF CIF 1763734.

Ethics Statement

Our work offers a method to augment the reviewer pool of large conferences with reviewers that are not considered by conventional reviewer recruiting methods. The small-scale experiment that we conducted in the ICML 2020 conference demonstrated that a combination of recruiting and mentoring mechanisms coupled with the reduced load results in reviews written by the new reviewers being a valuable addition to reviews written by the main pool of reviewers. This result hints that the scaled version of our procedure can help large AI and ML conferences to deal with rapid growth, thereby helping these fields to keep up the pace of the progress, positively impacting thousands of researchers. It is also known that exposing junior researchers to reviewing can also positively influence their own research perspective.

Despite promising results, we would like to underscore that peer review is a sensitive mechanism whose fallacies can have far-reaching consequences on the career trajectories of researchers. Hence, the results reported in this paper should be interpreted with utmost care. To emphasize this point, in addition to the caveats mentioned in the main text of the paper, we would like to make several other remarks:

- First, while we measured a number of metrics that were possible to measure and have been considered in the literature, we cannot exclude the possibility that EXPERIMENTAL reviewers may be worse than the main pool of reviewers in some other aspect not considered here.
- Second, it is possible that behavior of EXPERIMENTAL reviewers was affected by demand characteristics (McCambridge, De Bruin, and Witton 2012), that is, EXPERIMENTAL reviewers could hypothesize that we want them to perform better than reviewers from the main pool and hence they could adjust their behaviour to meet these perceived expectations.
- Finally, in extrapolating any results to other conferences, one should carefully consider any idiosyncracies of specific conferences.

All the aforementioned caveats coupled with sensitivity of the subject matter underscore the importance of a careful experimentation with the proposed procedure before its implementation in the routine review process.

References

Beygelzimer, A.; Fox, E.; dÁlche Buc, F.; and Larochelle, H. 2019. What we learned from NeurIPS 2019 data. https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2019-data-111ab996462c [Accessed: 9/7/2020].

Charlin, L.; and Zemel, R. S. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. URL http://www.cs.toronto.edu/~zemel/documents/tpms.pdf. [Accessed: 3/20/2021].

Feldmann, A. 2005. Experiences from the Sigcomm 2005 European shadow PC experiment. *ACM SIGCOMM Computer Communication Review* 35(3): 97–102.

Fisher, R. A. 1935. *The design of experiments.* Oxford, England: Oliver & Boyd.

Gao, Y.; Eger, S.; Kuznetsov, I.; Gurevych, I.; and Miyao, Y. 2019. Does My Rebuttal Matter? Insights from a Major NLP Conference. *CoRR* abs/1903.11367. URL http://arxiv.org/abs/1903.11367.

Garisto, D. 2019. Diversifying peer review by adding junior scientists https://www.natureindex.com/news-blog/diversifying-peer-review-by-adding-junior-scientists [Accessed: 9/7/2020].

Gurevych, I.; Miyao, Y.; and Cardie, C. 2018. A Report on the Review Process of ACL 2018. https://acl2018.org/2018/05/19/how-decisions-made/ [Accessed: 9/7/2020].

Kerzendorf, W. E.; Patat, F.; Bordelon, D.; van de Ven, G.; and Pritchard, T. A. 2020. Distributed peer review enhanced with natural language processing and machine learning. *arXiv preprint arXiv:2004.04165* .

Köhler, T.; González-Morales, M. G.; Banks, G. C.; O'Boyle, E. H.; Allen, J. A.; Sinha, R.; Woo, S. E.; and Gulick, L. M. 2020. Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology* 13(1): 1–27.

Kotturi, Y.; Kahng, A.; Procaccia, A. D.; and Kulkarni, C. 2020. HirePeer: Impartial Peer-Assessed Hiring at Scale in Expert Crowdsourcing Markets. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20. AAAI Press.

Kurokawa, D.; Lev, O.; Morgenstern, J.; and Procaccia, A. D. 2015. Impartial Peer Review. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 582–588. AAAI Press. ISBN 9781577357384.

Langford, J. 2018. When the bubble bursts.... https://hunch.net/?p=9604328 [Accessed: 9/7/2020].

Levy, P.; and Sarne, D. 2018. Understanding Over Participation in Simple Contests. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

Lian, J. W.; Mattei, N.; Noble, R.; and Walsh, T. 2018. The Conference Paper Assignment Problem: Using Order

Weighted Averages to Assign Indivisible Goods. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

McCambridge, J.; De Bruin, M.; and Witton, J. 2012. The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS one* 7(6): e39116.

McCook, A. 2006. Is peer review broken? Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. What's wrong with peer review? *The scientist* 20(2): 26–35.

Mogul, J. C. 2013. Towards More Constructive Reviewing of SIGCOMM Papers. *SIGCOMM Comput. Commun. Rev.* 43(3): 90–94. ISSN 0146-4833.

Patat, F.; Kerzendorf, W.; Bordelon, D.; Van de Ven, G.; and Pritchard, T. 2019. The Distributed Peer Review Experiment. *The Messenger* 177: 3–13. doi:10.18727/0722-6691/5147.

Picciotto, M. 2018. New Reviewer Mentoring Program. *Journal of Neuroscience* 38(3): 511–511. ISSN 0270-6474. doi:10.1523/JNEUROSCI.3653-17.2017. URL https://www.jneurosci.org/content/38/3/511.

Sculley, D.; Snoek, J.; and Wiltschko, A. B. 2019. Avoiding a Tragedy of the Commons in the Peer Review Process. *CoRR* abs/1901.06246. URL http://arxiv.org/abs/1901.06246.

Shah, N. 2019a. Double Decker Peer Review. Research on Research blog. https://researchonresearch.blog/2019/02/23/double-decker-peer-review/. [Accessed: 3/20/2021].

Shah, N. B. 2019b. Principled Methods to Improve Peer Review. URL http://www.cs.cmu.edu/~nihars/publications/survey_peerreview_niharshah.pdf. [Accessed: 3/20/2021].

Shah, N. B.; Tabibian, B.; Muandet, K.; Guyon, I.; and Von Luxburg, U. 2018. Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research* 19(1): 1913–1946.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. In *NeurIPS*.

Stelmakh, I.; Shah, N. B.; and Singh, A. 2018. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. *arXiv preprint arXiv:1806.06237* .

Stelmakh, I.; Shah, N. B.; Singh, A.; and Daumé III, H. 2020. A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences. *arXiv preprint arXiv:2011.15050* .

Tomiyama, A. J. 2007. Getting Involved in the Peer Review Process. https://www.apa.org/science/about/psa/2007/06/student-council [Accessed: 9/7/2020].

Toor, R. 2009. Reading Like a Graduate Student. https://www.chronicle.com/article/Reading-Like-a-Graduate/47922 [Accessed: 9/7/2020].

Wilson, E. B. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* 22(158): 209–212. ISSN 01621459. URL http://www.jstor.org/stable/2276774.

Xu, Y.; Zhao, H.; Shi, X.; and Shah, N. 2019. On Strategyproof Conference Review. In *IJCAI*.