What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review

Terne Sasha Thorn Jakobsen[1], Anna Rogers[1,2]
[1] Copenhagen Center for Social Data Science, University of Copenhagen
[2] RIKEN Center for Computational Science
[1]terne.thorn@sodas.ku.dk, [2]arogers@sodas.ku.dk

Abstract

Both scientific progress and individual researcher careers depend on the quality of peer review, which in turn depends on paper-reviewer matching. Surprisingly, this problem has been mostly approached as an automated recommendation problem rather than as a matter where different stakeholders (area chairs, reviewers, authors) have accumulated experience worth taking into account. We present the results of the first survey of the NLP community, identifying common issues and perspectives on what factors should be considered by paper-reviewer matching systems. This study contributes actionable recommendations for improving future NLP conferences, and desiderata for interpretable peer review assignments.

1 Introduction

Peer review is increasingly coming under criticism for its arbitrariness. Two NeurIPS experiments (Price, 2014; Cortes and Lawrence, 2021; Beygelzimer et al., 2021) have shown that the reviewers are good at identifying papers that are clearly bad, but the agreement on the "good" papers appears to be close to random. Among the likely reasons for that are cognitive and social biases of NLP reviewers (see overview by Rogers and Augenstein, 2020), fundamental disagreements in such an interdisciplinary field as NLP, and acceptance rates that are kept low[1] irrespective of the ratio of high-quality submissions.

Such arbitrariness leads to understandable frustration on the part of the authors whose jobs and graduation depend on publications, and it also means lost time and opportunities (Aczel et al., 2021; Gordon and Poulin, 2009) for science overall. Reviews written by someone who does not have the requisite expertise, or does not even consider the given type of research as a contribution, it is a loss for all parties: the authors do not get the intellectual exchange that could improve their projects and ideas, and reviewers simply lose valuable time without learning something they could use. It is also a loss for the field overall: less popular topics could be systematically disadvantaged, leading to ossification of the field (Chu and Evans, 2021).

[1]https://twitter.com/tomgoldsteincs/status/1388156022112624644

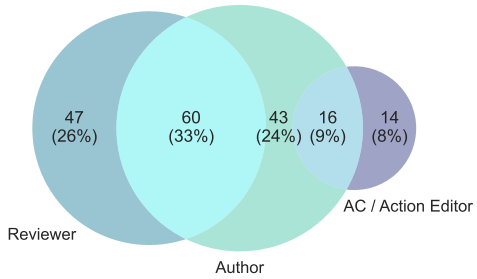


Figure 1: Overview of all respondents and overlap of their roles for their last experience at NLP venues.

This paper contributes a snapshot of this problem in NLP venues, based on a survey of authors, reviewers and area chairs (ACs). We collected 180 responses, which is is comparable to the volume of feedback collected for implementing the ACL Rolling Review (ARR). The overall distribution of respondents' roles is shown in fig. 1. We present the commonly reported issues and community preferences for different paper assignment workflows (section 4). We derive actionable recommendations to how peer review in NLP could be improved (section 5), discuss the limitations of survey methodology (section 6.2), and conclude with desiderata for interpretable peer review assignments (section 6.3).

2 Background: Peer Review in NLP

Paper-reviewer assignments are matches between submissions to conferences or journals and their available pool of reviewers, taking into account the potential conflicts of interest (COI) and reviewer assignment quotas.

Among the systems used in recent NLP conferences, the Softconf matching algorithm takes into account bidding, quotas, and manual assignments, and randomly assigns the remaining papers as evenly as possible[2]. NAACL and ACL 2021 used SoftConf, but also provided their ACs with affinity scores produced by a "paraphrastic similarity" system based on an LSTM encoder, which is trained on Semantic Scholar abstracts (Wieting et al., 2019; Neubig et al., 2021). Affinity scores are scores indicating how well a given submission matches a given reviewer. They are typically computed as the similarity (e.g. cosine similarity) between the embeddings of certain information about the submission and the reviewer's publication history (e.g. abstracts and titles).

ARR switched to OpenReview and currently uses[3] their SPECTER-MFR system (OpenReview, 2021) which is based on SPECTER (Cohan et al., 2020) and MFR embeddings (Chang and McCallum, 2021) for computing affinity scores. The assignments are then made with the MinMax matching algorithm[4].

The problem of paper-reviewer assignment is by itself an active area of research (see overview of key issues for CS conferences by Shah (2022)). There are many proposals for paper-reviewer assignment systems (Hartvigsen et al., 1999; Wang et al., 2010; Li and Watanabe, 2013, inter alia), some of which also consider the problem of "fair" assignments (Long et al., 2013; Stelmakh et al., 2019). Such studies tend to be hypothesis-driven: they make an assumption about what criteria should be taken into account, design a system and evaluate it. To the best of our knowledge, ours is the first study in the field to address the opposite question: what criteria should be taken into account, given the diversity of perspectives in an interdisciplinary field? We take that question to the community.

3 Methodology: survey structure and distribution

We developed three separate surveys for the main groups of stakeholders in the peer review process: authors, reviewers and ACs.

They follow the same basic structure: consent to participation (see Impact Statement), background information, questions on most recent experiences in the role which the survey pertains to, and how the respondents believe paper-reviewer matching should be performed. Most questions are asked to respondents in all three roles, reformulated to match their different perspectives.

The responses were collected late 2021 and all respondents are required to confirm that their most recent experience as an AC/reviewer/author is in 2019-2021. The full surveys and response data are publicly available[5].

Participant background. All surveys include questions on career status and the number of times the respondents have been ACs/reviewers/authors at NLP venues. We ask what venues they have experience with (as broad categories) and what types of contributions they make in their work.

Participant experience with peer review. We further ask the respondents a range of questions about their experience as AC/reviewer/author: how satisfied they are with the process, what issues they have experienced, what was the assignment load (ACs and reviewers), how paper-reviewer matching was done, how they would prefer it to be done, and which factors they believe to be important for paper-review matching. Most of the questions are multiple-choice, with addition of some open-ended questions where appropriate, so that respondents can elaborate their answers or add to the available options. Whenever possible, the question formulations were taken from the question bank of UK Data Service (Hyman et al., 2006). Attitude questions use a 5-point Likert scale.

Limited memory is an important concern in surveys (Sudman and Bradburn, 1973; Öztas Ayhan and Isiksal, 2005), and we cannot expect the respondents to accurately recall all their experience with peer review. To reduce memory recall errors, the survey focuses on the respondent's most recent experience, but they also have a chance to reflect on prior experience in open-ended questions, and to report whether they experienced certain issues at any time in their career.

Survey distribution. We distributed the surveys via three channels: by handing out flyers at EMNLP 2021, through mailing lists (ML-news, corpora list, Linguist list), and through Twitter with the hashtag #NLProc. Participation was voluntary,

[2] https://www.softconf.com/about/index.php/start/administration-view

[3] Source: personal communication with the ARR team.

[4] https://github.com/openreview/openreview-matcher

[5] https://github.com/terne/Paper-Reviewer-Matching-Surveys

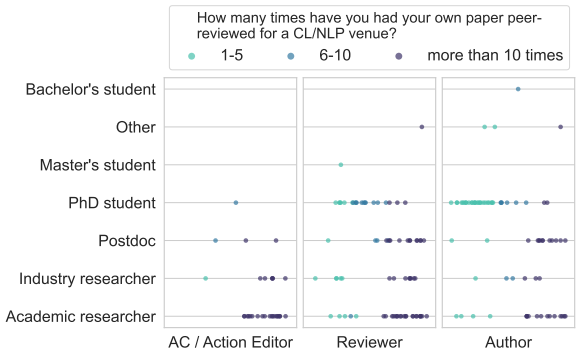


Figure 2: Career status of the respondents vs their experience receiving peer review. Numerical data is available in table 1 in the appendix.

with no incentives beyond potential utility of this study for improving NLP peer review.

Data validation. Given that links to surveys were distributed openly and that we did not ask for any identifiable information, the surveys needed to include other means of validation to ensure that the responses included in the analysis were from attentive, relevant individuals. Our approach for validating the data quality follows *satisficing theory* (Liu and Wronski, 2018), with the main safeguards being 1) the checking of response consistency, including a few "traps" where inconsistency or illogical responses can be exposed, and 2) the inclusion of open-ended questions.

73% ACs, 40% reviewers, 33% authors have provided at least one response to our open-ended questions, and we did not find any meaningless or incoherent comments not addressing the question. For consistency checks, all respondents were asked:

- How many times they have been an AC/reviewer/author. One of the options was "0", contradicting the earlier confirmation of experience in a given role.
- When was the last time they were an AC/reviewer/author. One of the options was "earlier than 2019", contradicting the earlier confirmation of peer review experience in 2019-2021.
- Whether they have performed the other roles. New authors may have not reviewed or AC-ed, but reviewers should also have been authors, and ACs should have experience with all roles.

4 Results

Overall we received 38 responses from ACs, 87 from reviewers and 81 from authors (206 in total). After removing 20 incomplete responses and 8 responses inconsistent with the "trap" questions, we report the results for 30 responses from ACs, 77 from reviewers and 73 from authors (180 in total).

4.1 Who are the respondents?

According to the past conference statistics, we could expect that many submissions would be primarily authored by the students, and reviewers are generally expected to be relatively senior, which should correspond to their going through peer review more often. We can use this expected pattern as an extra validation step for the survey responses.

Figure 2 shows that the responses are in line with this expected pattern. We received the most responses from academic researchers (62), PhD students (54), and postdocs (32). Most academic researchers and postdocs, but not PhD students, have had their work reviewed more than 10 times. At the same time 65% of the PhD students who served as reviewers went through peer review more than 5 times, as opposed to 24.2% of PhD students in the author role. Fewer industry than academic researchers responded to the survey. This could be related to the fact that a large part of the "academic" demographic are students – and in 2020-2021 the ACL membership among students was equal to or exceeding other demographics (Rasmussen, 2021).

4.2 Paper types

The next question is to see what kinds of research papers the respondents to our surveys have authored: engineering experiment, survey, position paper etc., according to the COLING taxonomy by Bender and Derczynski (2018). We expect that more senior researchers will have more experience

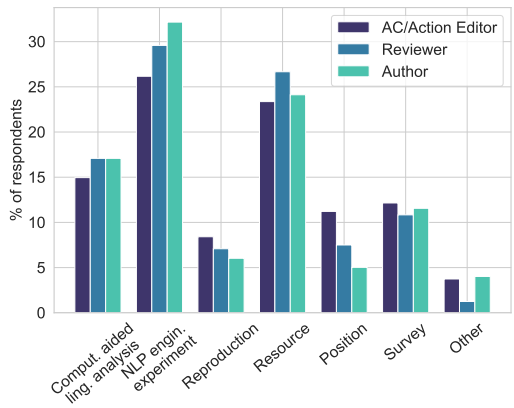


Figure 3: Types of research performed by respondents (multiple options could be selected).

with different types of work. Indeed, on average the authors have worked with 2.5 types of papers, vs. 3.0 for reviewers and 3.6 for ACs. The distribution is shown in fig. 3. The most respondents have authored engineering experiment papers (with the authors reporting the most such work).

Note that this only indicates whether the respondents to our surveys have or have not authored certain types of papers, rather than how many. In terms of volume, the engineering papers are a lot more prevalent: e.g. at ACL 2021 the "Machine learning" track had 332 submissions, vs 168 in the "Resources and evaluation" track (Xia et al., 2021).

4.3 What kinds of problems do people report?

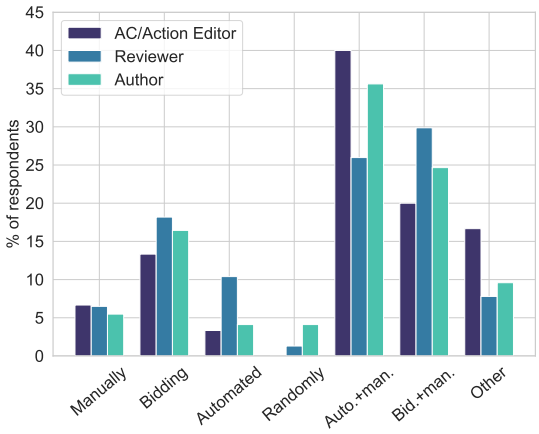
As with any voluntary feedback, our surveys were likely to receive more responses from people who had a grievance with the current process. Indeed, we find that only 6.7% of ACs, 20.5% of authors, and 22.1% of reviewers say that they have not had any issues in their last encounter with NLP venues.

The overall distribution for the types of problems reported by the authors, reviewers and ACs in their last and overall experience is shown in fig. 5. Given that at the time of this survey the ARR was recently deployed as the only ACL submission channel, we highlight the responses from the people for whom the most recent venue was ARR: 28% reviewers, 18% authors, 50% ACs.

The key takeaways are as follows:

- Two of the most frequent complaints of ACs (about 50% of the respondents) are insufficient information about reviewers and clunky interfaces;
- Many paper-reviewer mismatches (about 30%, if the report of the last experience is representative) are *avoidable*: they should have been clear from the reviewers' publication history;
- Over a third of the author respondents in their last submission (about 50% over all history) received reviews from reviewers lacking either expertise or interest, and that is supported by the reviewers' reports of being assigned papers that were mismatched on one of these dimensions;
- The authors report that many reviews (over a third in last submission, close to 50% overtime) are biased or shallow, which might be related to the above mismatches in expertise or interest;
- Two patterns are exclusive to ARR: insufficient time for ACs[6], and zero authors with no issues.

[6]ARR has since switched to 6-week cycles, which might help to address this issue (https://aclrollingreview.org/six-week-cycles/).



Topics mentioned in the open-ended comments
(See supplementary materials for full categorized comments)

ACs: bidding (2), similarity+manual (1), similarity+bidding+manual (5), keyword-based filtering + bidding (2), similarity (1), tracks (1), other info (2), ARR (1), interface (1)

Reviewers: manual (2), similarity + bidding (3), similarity+bidding+manual (3), keywords (1), keywords+similarity (1), tracks (2), tracks+bidding (1), other (4)

Authors: against similarity (2), similarity + bidding (2), similarity+bidding+manual (2), ARR (2), random (2)

Figure 4: *Which of the following options would you consider best for assigning reviewers to submissions?*

4.4 Knowledge of the workflow

Our next question is what methods NLP venues use to match submissions to reviewers, and to what extent the stakeholders (authors and reviewers) are aware of how it is done. We find that relatively few authors (23.3%) and reviewers (23.4%) know for sure what process was used, which begs for *more transparency in the conference process*. The ACs report that the most frequent case (37%) is a combination of automated and manual assignments. Interestingly, most reviewers believe that their assignments were automated (36%), and only (28%) believe they were automated+manual. See App. Figure 8 for full distribution.

5 The Ideal Process

5.1 Ideal workflow

When asked about what paper-assignment process they would prefer (given that fully manual matching is impractical for large conferences), most ACs and authors opted for automated+manual process, but for the reviewers this is the second preferred process (26%), with 30% opting for bidding + manual checks (see fig. 4). There was also relatively

Topics mentioned in the open-ended comments
(The full comments categorized by these topics can be found in the survey data repository)

Area chairs: interface issues (7), bad reviewers/reviews (5), workload issues (6), issues with ARR (4), lacking information on reviewers (4), communication issues between both systems and other human agents (4), lack of qualified reviewers in the pool (3), issues with meta-reviews (2), affinity score complaints (2), affinity score for finding reviewers the AC does not know personally (1), preference for manually recruited reviewers (1), papers assigned to ACs outside their area of expertise (1), too many declines (1), mismatch in goals of reviewers and authors (1), emergency reviews (1), bidding enabling bias (1).

Reviewers: choices forced by ACs (5), preference for bidding (4), areas of past expertise not currently of interest (4), lack of interest in the paper (3), methodological mismatch between generations of NLP researchers (3), mismatch in research methods (2), publication records as an unreliable indicator for assignments (1), mismatch in languages (1), time issues (1), reviewer bias (1)

Authors: reviewer expectation for a certain kind of research (6), inattentive reviews (5), short reviews (3), mismatch between the score and the text of the review (3), requests for irrelevant citations (2), confirmation bias (1), non-constructive criticism (1), shallow reviews (1), lack of reviewer competence (2), missing reviews (2), requests for irrelevant comparisons (1), "wild" estimates of impact (1), unannounced policy changes (1)

Figure 5: The issues with peer review process, reported by ACs, reviewers and authors, in their last (on the left) versus historical (on the right) experience with CL/NLP venues.

large support for pure bidding (13-18% of respondents in all roles), and cumulatively pure bidding and bidding with manual adjustments have as much or more support from all respondent categories than the automated matching + manual assignments.

The analysis of open-ended comments suggested that the respondents were aware that bidding is quite labor-intensive on the part of the reviewers. 5 ACs, 3 reviewers and 2 authors suggested using affinity scores to filter the papers on which bids would be requested, followed with manual checking. Another suggestion was keywords or more fine-grained areas/tracks, potentially as alternative to affinity scores for filtering down the list of papers to bid on. One AC suggested *"an extensive, but still finite, set of tags (e.g. an ACL-version of ACM CCS concepts, or FAccT's submission tags"*. One reviewer stressed that the keywords should be provided by the authors, to match what *they* perceive to be the salient aspects of the paper.

1 reviewer and 1 author suggested looking at whether the paper *cites* the potential reviewer[7], as this could be a good indicator for the reviewer's interest. 1 reviewer and 2 authors voiced support for some randomness in the assignments (given a track-level match): *"Bidding + some random assignment to ensure diversity in the matching. We don't want reviewers to review only papers they *want* to review. However these random assignments should be clearly indicated to all, and treated accordingly."*

5.2 Ideal assignment criteria

AC past experience. Figure 5 shows that one of the most common problems for the ACs is that they were not provided with enough information to facilitate the paper-reviewer matching. The follow-up question is what information they *are* provided with, and how useful they find it.

Figure 6 shows that the types of information with the highest utility information are links to reviewer profiles, bidding information, and affinity scores. But affinity scores are also the most controversial: it is the type of information that the most ACs find "not very useful" or "not useful at all" (20%).

Overall the results suggest that ACs are presented with little structured information about reviewers, and have to identify the information they need from a glance at the reviewers' publication record. Seniority, expertise, and reviewer history notes from other ACs are all reported to be useful, but they were never provided directly to many ACs.

An avenue for future research is offered by three types of information that the most ACs are not sure about, presumably because they are rarely provided: structured information about the methods that the reviewers were familiar with, the languages they spoke, and affinity score explanations. We will show below that there is much support for taking such methods into account. For the languages, this might be due to the "default" status of English (Bender, 2019). We hypothesize that providing this information would make it easier to provide better matches for papers on other languages, which would in turn encourage the authors to submit more such work. Affinity will be discussed in section 6.3.

Stakeholder preferences. We then asked the respondents what factors they believe are important for paper-reviewer assignments. Their answers are shown in fig. 7. The overall mean importance rankings (on scale 0-5) are as follows:

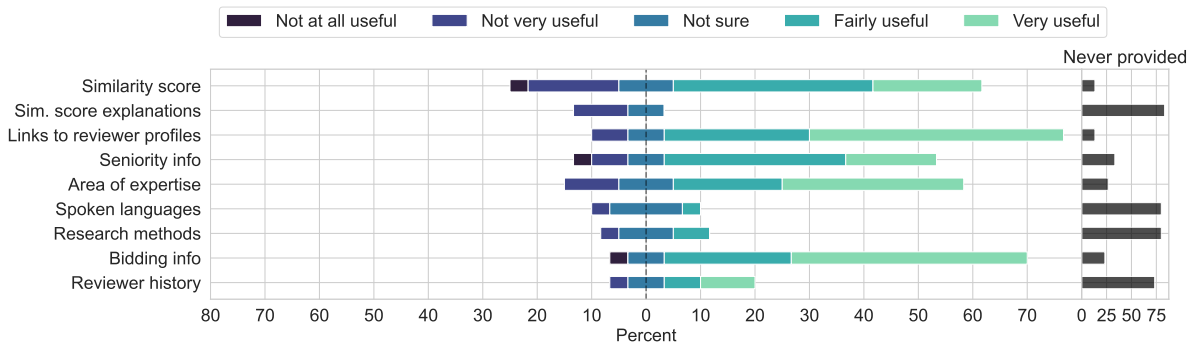
3.95 Reviewer has worked on the same task
3.85 Reviewer bid on the paper
3.72 Reviewer has worked with the same method
3.32 Reviewer has authored the same type of paper
3.11 AC knows & trusts the reviewer
2.81 Reviewer has worked with the same kind of data
1.99 The affinity score is high

The fact that affinity scores rank the least important for NLP researchers (who would know the most about them) is interesting, and perhaps related to the fact that evaluation of paper-reviewer matching systems remains an open problem, with little empirical evidence for how well our current systems really work. In the absence of such evidence, our results suggest that the respondents across all groups are not very positive about their experience with such systems. In the authors' personal experience, when the conference chairs provide automated affinity scores they caution the area chairs against fully relying on them, and urge to adjust the assignments manually.

Our data suggests that within groups of stakeholders the individual variation in importance of different factors is higher for some factors and stakeholders than others: e.g. ACs vary within 1 point on the importance of knowing the data, but only within 0.74 points on importance of knowing the tasks. This has implications for approaches who would rely on AC assignments as ground truth for automated assignment systems: they could end up modeling the annotator instead of the task (Geva et al., 2019). See App. table 2 for full data.

We then explored the question of whether the experience of having authored research of a certain type correlates with any changes in the attitude towards some of these paper-reviewer matching factors. For each pair of type of research and matching factor, we ran two-sided Fisher's Exact tests for all respondents who have authored (or not) the types of research and the importance they attached to different factors in paper-reviewer assignment (binning on less than moderately important and more than moderately important). For some pairs there were statistically significant differences: e.g. the respondents who have authored reproduction papers were significantly more likely to believe it important that the reviewer has worked with the same kind of data ($p = 0.004$), and respondents who authored position papers were significantly *less* likely

[7] We believe this is an interesting idea, but it could lead to authors strategically placing citations to maximize the chances of acceptance, or being punished for citing work that they may criticize or claim to improve upon.



Topics mentioned in the open-ended comments: reviewer history (2), number of assigned papers (1), being able to ask SACs for advice (1), reviewer affiliation (e.g. academic or industry) (1), correct area match for both ACs and reviewers (1).

Figure 6: The diverging bars shows the experienced utility of different kinds of information about reviewers that ACs may have been presented with to assist in manual checks of paper-reviewer matches. If the respondent had never been presented with the specific kind of information they chose "Never provided".

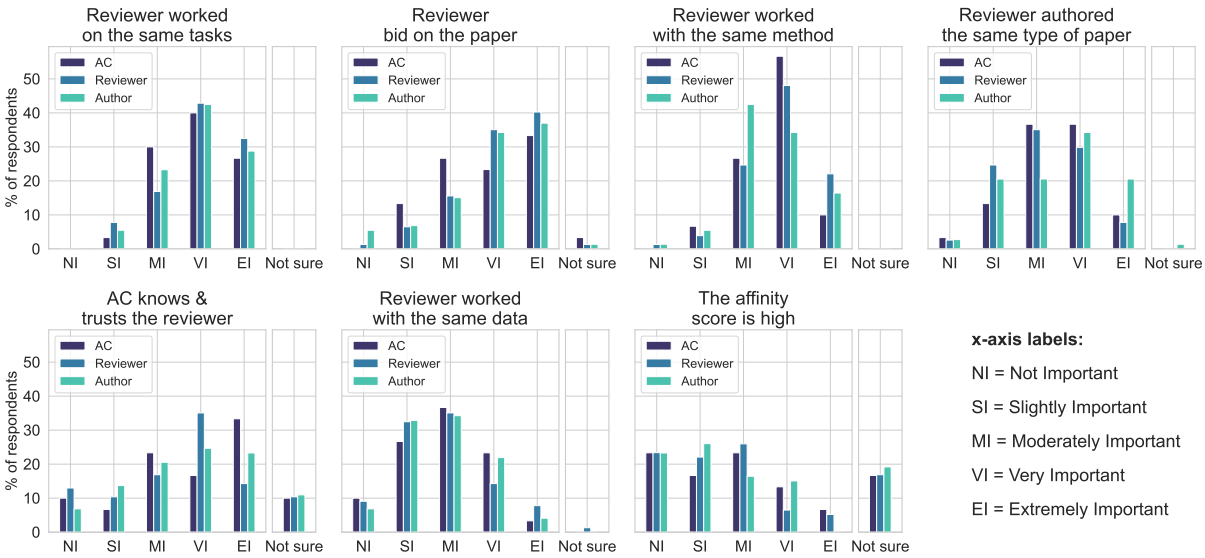


Figure 7: Question: *How important do you think the following factors are for a good paper-reviewer match?*

to believe a high automated affinity score is important ($p = 0.003$). See table 3 in the appendix for all p-values and more details on the tests. We note that the relationships are not necessarily causal.

We conclude that our sample does provide evidence (the first, to our knowledge) that researchers in interdisciplinary fields who perform different kinds of research may have differing preferences for what information should be taken into account for paper-reviewer assignments. If that effect is robust, it should be considered in assignment systems for interdisciplinary fields. We hope that this finding would be explored in a larger study, taking into account both the experience of authoring a given type of paper and how central that type of research is for a given researcher (a factor that we did not consider). Another direction for future work is exploring this question from the perspective of demographic characteristics and the type of institution the respondents work in. Should there be significant differences, more targeted assignments could be a powerful tool for diversifying the field.

5.3 Ideal workload

We asked our reviewer and AC respondents how many assignments they received at their most recent NLP venue, and what would be the optimal number (given a month to review, and a week for AC assignments). For ACs, the mean optimal number of assignments is 8.5±4.2 vs. 9.1±5.1 they received at the most recent venue, and for reviewers it is 2.8±1.0 vs. 3.3±1.8. Whether this is an

issue depends on how much time a given venue allows. The ARR reviewers have even less than a month, and they indicated preference for fewer assignments than they received (2.4±1.0 vs 3.3±1.9). See App. fig. 10 for data on other venues.

The lack of reviewers is a well-known problem. One of the possible causes is that many authors are students not yet ready to be reviewers. To investigate that, we asked the authors if they also reviewed for the venues where they last submitted a paper, and the reviewers and ACs - if they also submitted. If the core problem is that many authors are not qualified, we would expect more non-student authors to also be reviewers. Among all respondents there are 24% authors who submit to a venue but do not review there or help in some other role (fig. 1), but if we consider only non-student respondents that ratio is still 18% (see non-student role distribution in App. fig. 9). This suggests that *many qualified people do not review*.

6 Discussion

6.1 Reviewer interests

Our results suggest the lack of interest is one of the most common problems in paper-reviewer matching, for both authors and reviewers. The authors are aware of this problem and sometimes try to optimize for it by pursuing the "safe", popular topics. Unenthusiastic reviewers will likely produce shallow, heuristic-based reviews, essentially penalizing non-mainstream research. Both tendencies contribute to ossification of the field (Chu and Evans, 2021), and generally need to be minimized.

It is in the AC's interest to find interested reviewers, since that minimizes late reviews, but they need to know who finds what interesting. That is not as simple as a match by topic/methodology, clear from the publication record. Interests change not only gradually over time but also according to what is popular or *salient* at the given moment (Tversky and Kahneman, 1974; Dai et al., 2020), or even in seemingly irrational ways (e.g. by being sensitive to the framing of the problem) (Tversky and Kahneman, 1981). But although experience and knowledge may provide more stable descriptions of a reviewer, looking into dated publication records may be fundamentally counter-productive. According to one of our respondents: *"I prefer the conferences who offer bidding processes to select the papers to review... I am more enthusiastic to review the papers compared to conferences that assign papers based on what my interests were x years ago."*

Bidding however has its own set of problems, including the practical impossibility to elicit all preferences over a big set of papers, the possibility of collusion rings (Littman, 2021), and, as one of our respondents put it, "*biases towards/against certain paper types when bidding is enabled*". But these problems potentially have solutions: there is work on detecting collusion rings (Boehmer et al., 2022), and several respondents suggested that bidding could be facilitated by subsampling with either keyword- or affinity-score-based approaches.

We support some of our respondents' recommendation for a combination of interest-based and non-interest-based (within a matching area) assignments, with the latter clearly marked as such for ACs and reviewers, and separate playbooks for the two cases. The reviewer training programs should aim to develop the expectation that peer review is something that combines utility and exploration.

6.2 Limitations

We readily acknowledge that, like with any surveys with voluntary participation, our sample of respondents may not be representative of the field overall, since the people who have had issues with peer review system are more incentivized to respond. However, precisely for that reason this methodology can be used to learn about the commonly reported types of problems, which was our goal. Our response rate turned out to be comparable to the response rate of the official ACL survey soliciting feedback on its peer review reform proposal (Neubig, 2020), which received 199 responses.

It is an open problem how future conferences could systematically improve, if they cannot rely on surveys to at least reliably estimate at what scale an issue occurs. Asking about satisfaction with reviews does not seem to produce reliable results (Daumé III, 2015; Cardie et al.). Our survey included a question about satisfaction with the paper-reviewer matching, and whether the most recent experience was better or worse than on average. Both reviewers and authors were more satisfied than dissatisfied, and considered the recent experience better than on average, despite reporting so many issues (see App. fig. 11 for the distribution).

6.3 Interpretable Paper-Reviewer Matching: Problem Formulation

There already are many proposed solutions for paper-reviewer matching (see section 2), but their evaluation is the more difficult problem. The obvious approach would be to use bidding information or real assignments made by ACs as ground truth, but this data is typically not shared to protect reviewer anonymity. It would also provide a very noisy signal not just due to different assignment strategies between ACs, but also different quality of assignments depending on how much time they have on a given day. Both ACs and bidding reviewers are also likely[8] to favor top-listed candidates. And, as our findings suggest, the optimal assignment strategies in an interdisciplinary field might genuinely vary between different types of papers and tracks. A system unaware of that might systematically disadvantage whole research agendas.

Given that even the human experts cannot tell what the best possible assignments are, we propose to reformulate the problem as *interpretable paper-reviewer matching*. That problem is *not* the same as the problem of faithfully explaining why a given paper-reviewer matching system produced a certain score, for which we have numerous interpretability techniques (Søgaard, 2021). The AC goal is fundamentally different: not to understand the system, but to quickly find the information that the AC[9] considers relevant for making the best possible match. Therefore *the task of interpretable paper-reviewer matching is rather to help to identify the information that the stakeholders wish the decisions to be based on, and to provide that information as justification for the decisions*.

7 Conclusion

We present the results of the first survey on paper-reviewer assignment from the perspective of three groups of stakeholders in the NLP community: authors, reviewers, and ACs. The results point at a host of issues, some immediately actionable (e.g. providing the ACs with better information), some normative (e.g. different kinds of research may need different assignment strategies), and some open (e.g. how do we evaluate the effect of any changes to peer review process?) A big issue for both authors and reviewers is mismatches due to lack of interest, which is in tension with explorative aspects of peer review. We recommend to address this issue with a combination of assignments based on bidding and random matches within area, backed up by reviewer training.

[8] Position bias is well documented in search & recommendation systems (Craswell et al., 2008; Collins et al., 2018).

[9] Or the program chairs, should the conference aim to have consistent policies for all ACs.

Acknowledgments

Many thanks to Marzena Karpinska, Friedolin Merhout, and the anonymous reviewers for their insightful comments. We would also like to thank all our survey respondents, without whom this study would not have been possible.

Impact Statement

Broader impact. The study identifies types of information that could be used to provide better paper-reviewer matches. Used strategically by a conference, it could be a powerful tool for diversifying the field, by helping the non-mainstream papers find the reviewers more open to them. By the same token, if the entity organizing the review process aimed for suppressing such research, deprioritising this information could harm such papers. Our proposal of interpretable paper-reviewer assignments would mitigate this potential risk by requiring the organizers to disclose their rationale for any given match.

Personal data. The surveys are designed to not solicit any personally identifiable information (including comments about individual peer review cases in the past conferences), or demographic information about participants.

Potential risks. The respondents are participants in anonymous peer review process, and as such being tracked back to individual peer review cases could expose them to retaliation. The survey therefore did not solicit information about specific venues (only broader categories such as "*ACL conferences"), and we manually verified that the open-ended comments also do not contain references to specific cases. We thus foresee no potential risks from deanonymization of the respondents.

Informed consent. The respondents are informed about the organizers and the objective of the study: to identity current practises of paper-reviewer assignments in CL/NLP conferences and ways in which this process can be improved. Responses are anonymous and respondents consent

to the use and sharing of their responses for research purposes. Respondents must give consent to continue the survey.

Intended use. The survey data and forms will be made publicly available for research purposes.

Institutional approval. The study was approved by the Research Ethics Committee at the authors' institution.

References

Balazs Aczel, Barnabas Szaszi, and Alex O. Holcombe. 2021. A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1):14.

Emily M. Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.

Emily M. Bender and Leon Derczynski. 2018. Paper Types.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. The NeurIPS 2021 Consistency Experiment.

Niclas Boehmer, Robert Bredereck, and André Nichterlein. 2022. Combating Collusion Rings is Hard but Possible.

Claire Cardie, Iryna Gurevych, and Yusuke Miyao. ACL 2018: A Report on the Review Process of ACL 2018.

Haw-Shiuan Chang and Andrew McCallum. 2021. Cold-start paper recommendation using multi-facet embedding.

Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. A Study of Position Bias in Digital Library Recommender Systems. *arXiv:1802.06565 [cs]*.

Corinna Cortes and Neil D. Lawrence. 2021. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. *arXiv:2109.09774 [cs]*.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94, New York, NY, USA. Association for Computing Machinery.

Jane Dai, Jeremy Cone, and Jeff Moher. 2020. Perceptual salience influences food choices independently of health and taste preferences. *Cognitive research: principles and implications*, 5(1):2–13.

Hal Daumé III. 2015. Some NAACL 2013 statistics on author response, review quality, etc.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Richard Gordon and Bryan J. Poulin. 2009. Cost of the NSERC Science Grant Peer Review System Exceeds the Cost of Giving Every Qualified Researcher a Baseline Grant. *Accountability in Research*, 16(1):13–40.

David Hartvigsen, Jerry C. Wei, and Richard Czuchlewski. 1999. The Conference Paper-Reviewer Assignment Problem*. *Decision Sciences*, 30(3):865–876.

Laura Hyman, Julie Lamb, and Martin Bulmer. 2006. The use of pre-existing survey questions: Implications for data quality. In *Proceedings of the European Conference on Quality in Survey Statistics*, pages 1–8. Cardiff Wales, UK.

Xinlian Li and Toyohide Watanabe. 2013. Automatic Paper-to-reviewer Assignment, based on the Matching Degree of the Reviewers. *Procedia Computer Science*, 22:633–642.

Michael L. Littman. 2021. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44.

Mingnan Liu and Laura Wronski. 2018. Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, 60(1):32–49.

Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. 2013. On Good and Fair Paper-Reviewer Assignment. In *2013 IEEE 13th International Conference on Data Mining*, pages 1145–1150.

Graham Neubig. 2020. ACL Rolling Review Proposal.

Graham Neubig, John Wieting, Arya McCarthy, Amanda Stent, Natalie Schluter, and Trevor Cohn. 2021. ACL Reviewer Matching Code. Association for Computational Linguistics.

OpenReview. 2021. Paper-reviewer affinity modeling for OpenReview.

Eric Price. 2014. The NIPS experiment.

Priscilla Rasmussen. 2021. Acl membership report 2012-2021 summer statistics. *ACL Admin Wiki*.

Anna Rogers and Isabelle Augenstein. 2020. What Can We Do to Improve Peer Review in NLP? In *Findings of EMNLP*, pages 1256–1262, Online. Association for Computational Linguistics.

Nihar B Shah. 2022. An overview of challenges, experiments, and computational solutions in peer review (extended version). *Communications of the ACM*.

Anders Søgaard. 2021. Explainable Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 14(3):1–123.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pages 828–856.

Seymour Sudman and Norman M. Bradburn. 1973. Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, 68(344):805–815.

A Tversky and D Kahneman. 1981. The framing of decisions and the psychology of choice. *Science (American Association for the Advancement of Science)*, 211(4481):453–458.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Fan Wang, Ning Shi, and Ben Chen. 2010. A comprehensive survey of the reviewer assignment problem. *International Journal of Information Technology & Decision Making*, 09(04):645–668.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

Fei Xia, Wenjie Li, and Roberto Navigli. 2021. ACL-IJCNLP 2021 Program Chair Report.

H. Öztas Ayhan and Semih Isiksal. 2005. Memory recall errors in retrospective surveys: A reverse record check study. *Quality and Quantity*, 38(5):475–493. Copyright - Kluwer Academic Publishers 2004; Last updated - 2021-09-09.

A Appendix

In this appendix we introduce supplementary figures and tables.

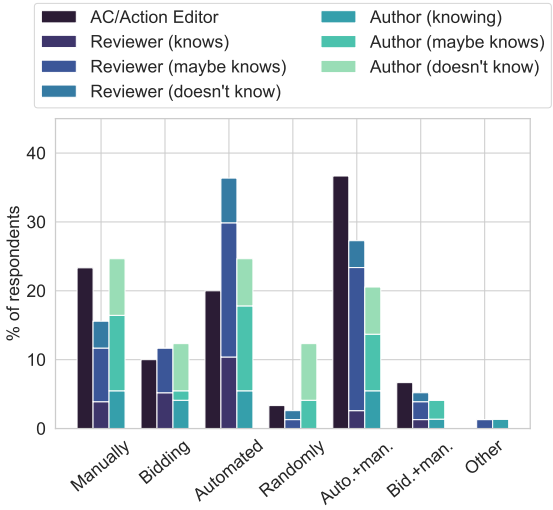


Figure 8: We ask reviewers and authors whether they know for certain, or maybe knows, or do not know how the paper-reviewer matching was done for their last CL/NLP venue. We then ask both reviewers, authors and ACs what they believe (or knows in the case of some) was the process for this venue The guesses, and knowledge herof, are much different from *best* options in fig. 4, discussed in section 5.

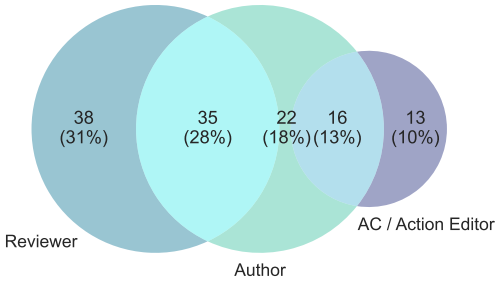
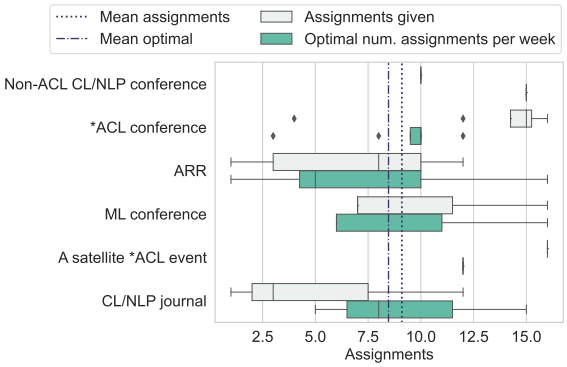
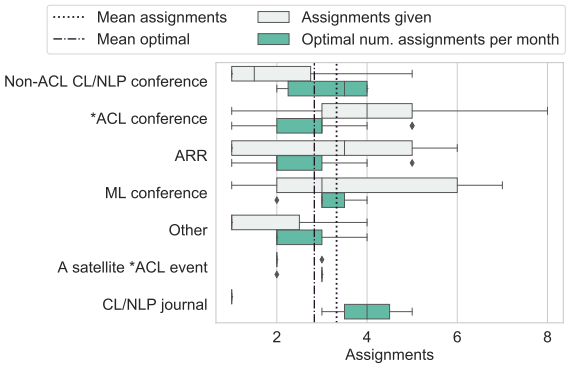


Figure 9: Distribution of **non-students** in the three roles, with overlap derived from asking the question *Did you also serve as a reviewer/author?* for their last CL/NLP venue.

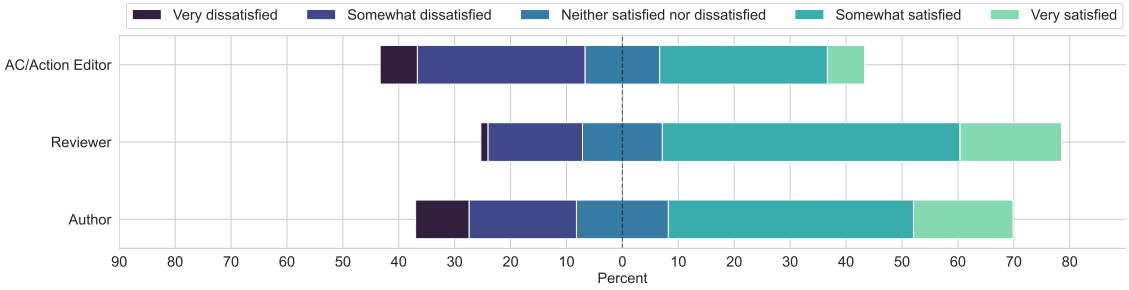


(a) AC/Action Editors

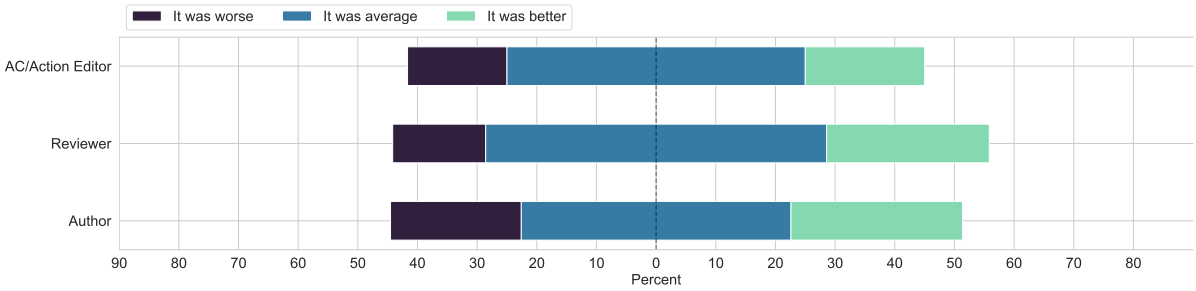


(b) Reviewers

Figure 10: The boxplots shows the number of assignments given and optimal for a) ACs and b) Reviewers, discussed in section 5.3. Number of given assignments are reported for the venue in which the respondent last served as AC/reviewer, and optimal number of assignments are reported for time periods one week for ACs and one month for reviewers. Mean given and optimal number of assignments, across all respondents/venues, are shown with vertical striped lines.



(a) Reported satisfaction with most recent experience. AC / Action Editor question: *(...) how satisfied were you with the support provided to you to improve the paper-reviewer matching?* Reviewer question: *(...) How satisfied were you with the paper-reviewer matching?* Author question: *(...) How satisfied were you with the amount of constructive criticism in the reviews you received?*



(b) Question: *would you say your most recent experience with paper-reviewer matching/paper assignment(s)/set of reviews described above, was better or worse than on average?*

Figure 11: We ask all respondents general questions about their satisfaction with their last experience as an AC/reviewer/author and their overall satisfaction in this role. We discuss these results in the limitations section 6.2.

	Area Chair			Reviewer			Author		
	1-5	6-10	>10	1-5	6-10	>10	1-5	6-10	>10
Bachelor's student	0	0	0	0	0	0	0	1	0
Master's student	0	0	0	1	0	0	0	0	0
PhD student	0	1	0	7	10	3	25	6	2
Other	0	0	0	0	0	1	2	0	1
Postdoc	0	1	2	1	3	12	2	0	11
Academic researcher	0	0	19	4	1	21	4	0	13
Industry researcher	1	0	6	5	0	8	1	2	3
Total	1	2	27	18	14	45	34	9	30

Table 1: This table shows the count of respondents from each role (AC/reviewer/author) reporting one of 7 career statuses and an amount of times having had their own papers reviewed. The numbers reflects those plotted in fig. 2, section 4.1.

	Tasks	Bidding	Method	Type of paper	Trust	Data	Affinity Score
AC / Action Editor	3.90±0.83	3.67±1.25	3.70±0.74	3.37±0.95	3.27±1.67	2.83±1.00	2.13±1.50
Reviewer	4.00±0.90	4.03±1.07	3.86±0.85	3.16±0.97	2.96±1.57	2.75±1.09	1.97±1.38
Author	3.95±0.86	3.86±1.22	3.59±0.87	3.45±1.18	3.11±1.60	2.84±0.98	1.85±1.32
Grand mean	3.95	3.85	3.72	3.32	3.11	2.81	1.99

Table 2: Mean importance with 0=Not sure, 1=Not important, 2=Slightly important, 3=Moderately important, 4=Very important and 5=Extremely important, for the seven paper-reviewer matching factors shown in fig. 7. Removing "Not sure" does not change the overall ranking. The grand mean is the unweighted mean of ACs', reviewers' and authors' mean scores. The mean absolute difference is greatest between ACs and reviewers (0.20) and smallest between ACs and authors (0.12), while between reviewers and authors it is 0.16. These results are discussed in section 5.2 under "Stakeholder preferences".

	Tasks	Bidding	Method	Type of Paper	Trust	Data	Affinity Score
Computationally-aided linguistic analysis	1.000	0.624	0.205	0.089	0.699	**0.035**$^{<}$	0.654
NLP engineering experiment paper	1.000	0.716	**0.038**$^{>}$	0.135	0.270	0.772	0.699
Reproduction paper	0.457	0.766	0.055	0.601	1.000	**0.004**$^{>}$	0.562
Resource paper	0.728	0.433	0.727	0.162	0.818	1.000	0.616
Position paper	0.222	0.766	0.433	0.135	0.236	0.493	**0.003**$^{<}$
Survey paper	1.000	0.186	0.738	0.420	1.000	0.840	0.480
Other	0.601	0.052	1.000	**0.019**$^{<}$	0.733	0.202	0.063

Table 3: P-values of two-sided Fisher Exact tests, discussed in section 5.2. For each contribution type, we test the null hypothesis that there is no difference in whether respondents find a paper-match factor (from fig. 7) more than or less than moderately important, depending on whether or not individuals have worked on the specific types of papers (contribution types). For each contribution type i and paper-match factor j, a 2×2 contingency table is made with the counts of a) respondents having worked with type i and finding factor j *less* than moderately important, b) having worked with type i and finding factor j *more* than moderately important, c) having *not* worked with i and finding j *less* than moderately important, d) having *not* worked with i and finding j *more* than moderately important. The p-values reflect the probability of observing the given counts or something more imbalanced between those having and not having worked on type i. Significant p-values, $p < 0.05$, are in bold, and for these, superscript $>$ denotes that respondents having worked with i believe factor j is *more* than moderately important, and the superscript $<$ denotes the opposite.