Natural Language Engineering (2020), **26**, pp. 245–257
doi:10.1017/S1351324920000030

CAMBRIDGE
UNIVERSITY PRESS

Emerging trends: Reviewing the reviewers (again)

Kenneth Ward Church

Baidu, Sunnyvale, CA 94089, USA
E-mail: KennethChurch@baidu.com

(Received 28 January 2020; accepted 29 January 2020)

Abstract
The ACL-2019 Business meeting ended with a discussion of reviewing. Conferences are experiencing a success catastrophe. They are becoming bigger and bigger, which is not only a sign of success but also a challenge (for reviewing and more). Various proposals for reducing submissions were discussed at the Business meeting. IMHO, the problem is not so much too many submissions, but rather, random reviewing. We cannot afford to do reviewing as badly as we do (because that leads to even more submissions). Negative feedback loops are effective. The reviewing process will improve over time if reviewers teach authors how to write better submissions, and authors teach reviewers how to write more constructive reviews. If you have received a not-ok (unhelpful/offensive) review, please help program committees improve by sharing your not-ok reviews on social media.

Keywords: Reviewing; Growth; Randomness; Scale

1. Reviewing is broken

This paper is a follow-up on “Reviewing the Reviewers” (Church 2005), an opinion piece I wrote 15 years ago for “The Last Words” section of computational linguistics. “The Last Words” was intended to be controversial.

At the time, it was controversial to criticize reviewing. Reviewing was a sacred cow that one should not question. People had more confidence in the process than it deserved. Even though we all knew better, especially those of us who have served as program chairs and area chairs, it was convenient to believe in simple fairy tales like the reviewing-is-perfect myth. Reality, of course, is an inconvenient truth.

> Pleasant fairy tales like the reviewing-is-perfect myth make people comfortable. It is hard to believe in fairy tales when too many people know the facts. And it’s hard to maintain plausible deniability when everyone knows what everyone knows (Church 2005).

It is not pretty how the sausage is made. There was one particularly memorable face-to-face program committee (PC) meeting about 20 years ago (back when the PC used to meet face-to-face in a physical room). In those days, we would invite the chairs of the student workshop to the PC meeting so they see what actually happens. (The student workshop is an effort by ACL to help students; it is a separate meeting run by students for students co-located with the main meeting.) The chairs of the student workshop were shocked when they saw how we started the meeting with a quick triage. In just a few seconds per paper, we cut the pile down to a manageable size by assigning submissions into: (1) easy accept, (2) easy reject, and (3) not-so-easy. The student chairs felt, quite understandably, that given how much work authors put into writing these papers, we should at least give them more than a few seconds, even if the decision was easy. On the other hand, those with more experience had more "appreciation" for the limitations of face-to-face PC meetings.

It was clear at the time that the process was buckling under the load of ever-increasing submissions. Even with all the corner-cutting that we were doing at the time, we were just barely getting the job done. Since then, of course, there are even more submissions than there were then.

1.1 Scale trumps precedent

As we all know, submissions have continued to increase over the last 20 years. In Church (2017), I estimated the number of publications was increasing by 2.4x per decade for ACL[a] and 1.7x per decade for PubMed.[b] These inflation rates were measured over decades for the ACL Anthology and centuries for PubMed. Data on submissions are harder to come by than data on publications, but I suspect that submissions are probably scaling at about the same rate, given that most venues try to maintain relatively constant acceptance rates, typically around 20%.

Reviewing processes tend to be based more on precedent than reality. It has been suggested that inflation is a relatively recent phenomenon, but I believe these exponential increases have been going on for a long time (decades/centuries). It is only recently that the community has begun to "appreciate" the need for processes that scale more effectively than what we have always done. Eventually, we will need to address reality. If the input increases exponentially, and downstream processes are not designed to scale effectively with more and more input, then at some point, the system will blow up. The current load is already stressing the system, and future loads will be even more challenging.

Increasing submissions have already forced various "improvements." When I published my first ACL paper in 1980, the meeting was small enough to fit in a university auditorium. There were no posters and no parallel sessions. It was possible for each member of the PC to read most of the submissions. These days, larger meetings necessitate compromises: for example, parallel sessions, posters, expensive conference centers, higher registration fees, larger and more hierarchical PCs, virtual PC meetings (as opposed to face-to-face PC meetings), slower decisions, and more mistakes.

The challenge is to come up with structures that scale even more effectively. It has been suggested that scale makes it impossible to do a good job with reviewing, but I am not convinced. We can all agree, however, that scale makes it impossible to do what we have always done since it is clear that many of those precedents do not scale. When push comes to shove, scale needs to be taken more seriously than precedent. Scale is a requirement, unlike precedent (merely a nice-to-have).

1.2 We are well past denial

These days, no one would deny that the process is broken. The process is so broken that I no longer have to say much about how broken it is, though that would not stop me from doing so. Ranting, of course, does not accomplish much, but it feels *so* good.

Before ranting too much myself, let me quote from a more careful (and thoughtful) survey on reviewing, which starts out by declaring the process far from perfect:

> If the current system of peer review were to undergo peer review, it would undoubtedly achieve a "revise and resubmit" decision. As Smith (2010) succinctly stated, "we have little or no evidence that peer review 'works,' but we have lots of evidence of its downside." There is further evidence to show that even the fundamental roles and responsibilities of editors, as those who manage peer review, has little consensus (Moher 2017), and that

[a] https://www.aclweb.org/anthology/.

[b] https://www.ncbi.nlm.nih.gov/pubmed/.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

> tensions exist between editors and reviewers in terms of congruence of their responsibilities Chauvin (2015). These dysfunctional issues should be deeply troubling to those who hold peer review in high regard as a "gold standard" (Tennant 2017).

The problem of bad reviewing is not new. We started EMNLP in the early 1990s because ACL reviewing was too conservative. They were rejecting great papers that were breaking new ground (in statistical methods).

ACL's low recall has been great for other conferences. The best of the rejects are very good, better than most of the accepted papers, and often strong contenders for the best paper award at EMNLP. I used to be surprised by the quality of these rejects, but after seeing so many great rejects over so many years, I am no longer surprised by anything. The practice of setting EMNLP's submission date immediately after ACL's notification date is a not-so-subtle hint: Please do something about the low recall (Church 2005).

Part of the reason for EMNLP's success was our reviewing. Not only were we more receptive to new ideas (in empiricism) and new blood (especially in Asia), but we did what we did relatively quickly. Speed was not an option. The dates were compressed by constraints imposed by the main ACL meeting. Given that EMNLP was typically co-located with ACL, and the submission date was determined by ACL's notification date, everything had to be done faster than ACL. As the cliche goes, necessity is the mother of invention. We figured out how to get the job done given these realities. As it turned out, the constraints worked out well for EMNLP. The last submission date of the season is a good place to be. Authors appreciate relatively quick decisions.

Unfortunately, these days, EMNLP reviewing is no longer much of a differentiator. EMNLP timing is no longer much of an advantage. Both conferences take much of a year to do what they do. And both conferences are equally conservative (and random); they are routinely rejecting the best work by the best people. We used to feel ashamed when a paper was rejected, but it is now so common for great work to be rejected that we often hear about rejected work in keynote talks.

1.3 Reviews should be constructive and helpful

When we review for NSF,[c] they encourage reviewers to teach authors how to be more successful in the future. NSF makes it relatively easy to accept a proposal. Positive reviews are a joy to write. If I like a proposal, I rarely find it hard to say why. And even if I cannot defend my position, the author is unlikely to push back on a positive grade, so I do not have to say much to the author.

If I really like a proposal, I direct my feedback more toward internal NSF processes (and less toward the author). A positive review should help higher level committees within NSF to come to the decision I hope they will come to, namely that they should fund this proposal, and not some other proposal from some other committee in some other field. If I like the proposal (and I like my own field), that should be a relatively easy argument to make (and not at all an unpleasant chore).

On the other hand, NSF makes it harder and more painful to write negative reviews. If I do not like a proposal, I am expected to teach the authors how to do better in the future. Sometimes it is easy to offer constructive advice but not often. Since most proposals will be rejected (by construction), reviewing for NSF tends to be an unpleasant chore (by construction).

No matter how much I dread reviewing for NSF, I have deep respect for the goal of teaching authors to do better in future. It might be painful in the short term, but there are clear long-term benefits for all parties: NSF, authors, reviewers, their future students, science, tax payers, etc. As the saying goes, "no pain, no gain." Reviewing is a teaching opportunity. NSF has a mission to teach.[d] It makes sense to encourage constructive reviews (even if they are not much fun to write).

[c] https://www.nsf.gov/.

[d] Other funding agencies have other missions. Slide 36 of Footnote 5 explicitly discourages reviewers from teaching authors how to do better in the future.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

1.4 Cheap thrill reviews

While I appreciate the merits of the NSF approach, there are times when I wish I could just vent. It would feel so good, especially after having read way too many no-hoper proposals, to say something totally unhelpful and inappropriate. Sometimes I would want to tell the author to give up and find another career. But that would be wrong. Venting might feel good, but it is selfish (and pointless).

It is ok for a review to agree or disagree on a technical matter, but we have all seen reviews that say something inappropriate like "give up" or "find another career." I have even seen reviews like: "no one from your country/school/group does good work."

We have probably all received reviews that are not ok. It hurts when our work is rejected, but it hurts even more when the reason for the rejection is not ok.

My most recent EMNLP submission was rejected with this remark: "I recommend reading several recent ACL or EMNLP papers prior to submitting, to get a sense of the conventions of the field." Maybe this reviewer was trying to be helpful but probably not. During the rebuttal period, I wanted to mention some of my experience (former president of ACL and co-creator of EMNLP), but could not see how to do that within the restrictions of the blind reviewing process.

My suggestion for how to improve reviewing is to hold the organization responsible for the quality of reviews. When I chaired EMNLP, I would "disappear" reviews that were unhelpful/offensive. No one needs to see a review that would embarrass the organization (but if it goes out the door on my watch, the error is my fault). The officer on deck is responsible for whatever happens on the ship. It is bad if he knows about it, and worse if he does not.[e]

Many organizations have established quality checks to address these concerns. As mentioned above, it is unlikely that NSF would send an offensive review out the door. The European Commission also takes these issues very seriously; see quality checks on slides 33–36 of this briefing.[f] Many journals have many rules and processes, though unfortunately, in a randomized controlled trial, it was found that reviewers often do not do what they are asked to do Chauvin (2015).

ACL has also thought deeply about such issues. They post many links such as this,[g] though these links focus more on advice to junior reviewers and less on quality checks and responsibility. When mistakes inevitably happen, there is a tendency to blame others (too many submissions, software limitations, etc.)[h]; the officer on deck ought to accept responsibility for whatever happens on his watch. We need to get past denial before we can think about constructive solutions such as quality checks.

1.5 Please help PCs improve

I bet that many of you have even more embarrassing reviews than my most recent EMNLP rejection. Rumor has it that a number of seminal papers (Penn Treebank Marcus, Santorini, and Marcinkiewicz 1993, Page Rank Page 1999, ELMO Peters 2018) were rejected, and some of those rejections were not ok. These papers have had considerable impact, in terms of both academic impact (citations) and commercial success ($$), compared to most papers in the literature.

PCs will improve the quality of their work if we give them more feedback. Negative feedback loops are more effective than the open loop process we currently have. If you have received a not-ok review, please help PCs improve by sharing your not-ok reviews on social media as I have done here.[i] The process will improve over time if reviewers teach authors how to write better submissions, and authors teach reviewers how to write more constructive reviews.

[e] https://www.youtube.com/watch?v=34ag4nkSh7Q.

[f] https://ec.europa.eu/info/sites/info/files/fet-open_ria_2018_-_web-briefing_for_res.pdf.

[g] https://acl2020.org/reviewers/.

[h] See comments in response to Facebook post in Footnote 8.

[i] https://www.facebook.com/kenneth.church.332/posts/10156806503257701.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

Reviewers should be encouraged to review the other reviews, both to improve the reviewing process and to reduce the number of not-ok reviews that go out the door. The officers on deck (area chairs and program chairs) should be particularly concerned about not-ok reviews, because they do not want too many mistakes to show up on social media on their watch.

1.6 Conservative reviewers considered harmful

Reviewers are conservative. They prefer boring incremental papers to papers that break new ground.

Some reviewers are conservative and some are really conservative and some are really really conservative. Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam). Precedents are good; novelty is bad (Church 2005).

2. Encouraging growth

2.1 Diversity: A few big groups and many small groups

The reviewing process discourages growth (contributions from new blood on new topics). About 20 years ago, I was at a face-to-face PC meeting. After a couple of days of hard work, we performed the usual spot checks for diversity. The initial checks went well enough. We checked for gender. That was ok. We checked for geography (by continent). That was also ok.

Since I happened to be interested in smoothing (of language models) at the time, I suggested an additional novel diversity check. Is there a bias in favor of countries with lots of submissions? To check this, we made a group of "small" countries, countries with five or fewer submissions. There were 42 papers from "minor Europe" (small countries in Europe). One hundred percent of these papers were rejected (on the first pass). Then, we made a group for "minor Asia" (small countries in Asia). One hundred percent of these papers were also rejected (on the first pass).

Some of these "small" countries were small (Singapore) and some were not (China). These papers turned out to be super-important for the future of the community. These days, we receive lots of papers from Singapore and China, but that might not have happened if we had not taken a second look at these papers.

Why were we about to reject 100% of these papers? Clustering of these papers identified two issues that reviewers do not like: (1) non-native English and (2) novel topics.

2.1.1 Non-native English

It is a shame that English is as important as it is. There was a time when I was an area chair and there were two papers in my pile from the same lab in Japan. Superficially, the two papers looked very similar. Both discussed similar topics using similar data. But one of them received the top average score and the other received the bottom average score. How could that be? Does average score have no meaning at all?

I gave the two papers to a colleague (without telling him which received which score). He came back after a few days and said, "oh, isn't this interesting,"[j] which is code for "I'm not going to tell you which paper should be accepted and which should not." But he did tell me that one of them had better content and the other had better English. As you have probably guessed by now, the one with better English scored better than the one with better content.

[j] My first boss, who was also Japanese, would say "oh, isn't that interesting" when I asked an inappropriate question. I learned that that was a polite way for him to avoid answering stupid questions. One time I was traveling with him in Japan and I asked him to teach me how to say "no" in Japanese. I thought I knew how to say "yes," but not "no." When Osamu said, "oh, isn't that interesting," I realized that not only was he not going to teach me how to say "no" (because it is rude), but I had no idea how little I knew. I thought I knew the Japanese equivalent of "yes," but actually, that is not really "yes." It can mean anything from "yes" to "I heard the question, and I'm stalling for time while I am thinking of a polite way to say 'no.' "

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

2.1.2 Novel topics

Reviewers are often looking for easy excuses to reject a paper without thinking too much. Non-native English is a lame excuse but not an uncommon one. I am even more concerned about novel topics. Reviewers do not like breaking new ground. Reviewers tend to be more receptive to topics that have been discussed ad nauseam. Reviewers do not know what to do with novel topics. They tend to abstain on such papers, but abstentions will kill a paper in a competitive conference. (Abstentions should not count, but they do.)

Conservative reviewing is considered harmful to the long-term future of the organization.[k] The future of the organization depends on new blood. We need to liberalize the process, or else.

2.2 Encouraging growth ≠ maximizing randomness

Over the years, lots of innovations have been introduced to improve the reviewing process in various ways (including diversity). One of these innovations is double-blind reviewing. I have never been a fan of double-blind reviewing because, as a senior citizen, I like to believe that experience is a good thing. That is, I believe, all other things being equal, one would expect highly respected authors with more experience to write better papers than less well-known authors with less experience. That said, in addition to the experience prior, there are many factors at work in the reviewing process, some fair and some unfair, some productive and some counter-productive, and some merely random:

1. Experience Prior: On average, authors with more experience (and strong reputations) write better papers than less famous authors with less experience.
2. Unfair Bias: Reviewers tend to favor friends over strangers.
3. Growth Opportunities: Reviewers tend to be biased away from opportunities for growing the organization, for example, new blood, new topics.
4. Randomness: Reviewers are already too random. Withholding useful information will make them even more random.

My problem with double-blind reviewing is that withholding information will not only increase fairness but also randomness. If we were really interested in fairness (and nothing else), one could simply flip coins.

Obviously, we care about more than just fairness. Experiments supporting double-blind reviewing need to establish that the treatment (double-blind reviewing) is not only more fair than the control (traditional reviewing) but also more effective. Does the treatment accept better papers (with more potential for growth) than the control, or is treatment merely more random than the control? Thus far, most experiments (Tomkins, Zhang, and Heavlin 2017; Stelmakh, Shah, and Singh 2019) have more to say about fairness than effectiveness.

Pragmatically, there is little that we can do at this point. Double-blind reviewing has become so well-established that we probably cannot change it now. But it is important, when thinking about changes to the system, to find scalable solutions that reduce randomness.

2.3 Greed is good; randomness is not

IMHO, if one wants to improve diversity, there are better ways to do that than increasing randomness. Consider, for example, the "minor Europe" and "minor Asia" story above. It was far more effective to explicitly target opportunities for growth, as we did, than to throw darts randomly. There were relatively few submissions from "minor Europe" and "minor Asia" (by construction), and therefore, a random process probably would not have been nearly as effective as what we did (check for biases against countries with fewer submissions).

[k] https://en.wikipedia.org/wiki/Considered_harmful.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

The argument for diversity should not merely stop at fairness. Diversity is in the best (long-term) interest of the organization. Reaching out to Singapore and China in the 1990s produced long-term benefits in the decades to come.

We should invest in diversity, not merely because it is the right thing to do, but also because it is in the long-term interest of the organization. Randomness is considered harmful for the long-term health of the organization. Greed is good[l] for the organization; randomness is not.

2.4 Too many submissions → randomness → even more submissions

Why is reviewing so random? It has been suggested that we cannot afford to do a good job because there are too many submissions. Conferences are experiencing a success catastrophe. They are becoming bigger and bigger, which is not only a sign of success but also a challenge (for reviewing and more).

Various proposals for reducing submissions were discussed at the ACL-2019 Business meeting. IMHO, the problem is not so much too many submissions, but rather, random reviewing. We cannot afford to do reviewing as badly as we do (because that leads to even more submissions).

It is said that people have more children when they do not expect their children to survive (Van de Walle 1986). So too, people submit more papers when reviewing is unpredictable. Authors have figured out that one can eventually succeed by submitting the same paper again and again. Conferences are responding with rules intended to discourage/prohibit multiple submissions (and re-submissions), but such rules are too easy to game (by splitting the paper up into a bunch of similar papers with just enough differences to get past the rules).

In short, while I agree that ever-increasing submissions will stress the reviewing system, I am too much of an optimist to give up on the system. It has to be possible to do a good job.

On the contrary, I believe we cannot afford to do a bad job. As mentioned above, the system is already suffering from a lack of confidence. We are now well past denial. Everyone knows the system is broken. I used to be able to guess which papers would be accepted and which will be rejected, but these days, I cannot do much better than flipping coins. Authors do not have the confidence that good work will be appreciated by the reviewers, and therefore, they are gaming the system by submitting more and more papers.

2.5 Scale is good

Scale should be our friend. As conferences become larger and larger, the law of large numbers should work in our favor. Larger conferences increase the demand for both reviewing (submissions) and supply (qualified reviewers). The challenge is to make scale work for us and not against us. The increase in demand (submissions) is all too apparent, but how do we identify the increase in supply (qualified reviewers)? And how do we convince these qualified reviewers to do their share of the work?

Let us start with metrics. Whatever you measure, you get. Can we estimate supply and demand, now and in the future? How many papers do we need to review, and how many qualified reviewers do we have? Who is doing their fair share and who is not? What is a fair share?

Some people (the usual suspects) do more than their fair share and others (free-loaders) do less. Rather than ask the usual suspects to do more than their fair share (which isn't fair to them or anyone else), I have found it more effective to ask them to name names. FOAF[m] (friend of a friend) is a great way to find new blood.

What about free-loaders? How do we convince more people to do their share of the work? Let's start with metrics. Whatever you measure, you get. Can we estimate supply and demand, now and

[l] https://www.youtube.com/watch?v=VVxYOQS6gg.

[m] https://en.wikipedia.org/wiki/FOAF_(ontology).

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

in the future? How many papers do we need to review, and how many qualified reviewers do we have? Who is doing their fair share and who isn't? What is a fair share?

2.6 You owe 15 reviews for each publication

Years ago, my co-author, the statistician Bill Gale (personal communication), argued that we (collectively) owed society 15 reviews for each paper that we published. That might sound like a lot, but his argument was very simple (and convincing). He assumed an acceptance rate of 20% and three reviews per paper. His 15 is simply 3/20%. We owe society not only for the three reviews for our paper but also for the three reviews for the 80% that were rejected (because we want the pool of reviewers to be more qualified than the pool of authors that submit papers).

Actually, his 15 underestimates the true burden for at least three reasons: (1) students, (2) transients, and (3) free-loading.

2.6.1 Students

He and I did not have that many graduate students (since we were working in industry), but if one of our interns should get a paper into a conference, then we owe society the 15 reviews for the student, since reviewers should be more qualified than students.

Some professors have lots of students. They often think they are doing more than enough reviewing, but in fact, they owe society 15 reviews for each paper published by their lab. I suspect that much of the problem with scale is caused by a few large (successful) labs that are publishing lots of papers, but failing to review 15 times as many papers as they publish (or allowing too many students to do too much reviewing).

2.6.2 Transients

Gale's analysis also assumed steady state. In fact, the field is growing rapidly, especially in some parts of the world. Growth is a good thing. We do not want to discourage that. But growth can create challenging short-term transients. Let us assume that growth in a new part of the world typically starts with students, and students are not qualified to do reviewing. Then, those of us with experience in other parts of the world will need to pick up some of the slack (at least in the short-term). So, if the number of submissions is growing by 2x overall, then those of us with experience may need to review 30 papers next year for each paper that we publish this year (at least until we identify experienced reviewers in the part of the world that is expanding).

2.6.3 Free-loaders

Gale's analysis assumed that everyone will contribute their fair share. Since that is obviously unrealistic, we need to impose a tax on everyone to cover the burden imposed by free-loaders. I do not know how many free-loaders there will be, but as the process scales up, it should be pretty easy to estimate how large the free-loading tax needs to be.

> . . . often a small minority of researchers [. . .] perform the vast majority of peer reviews (Fox *et al.* (2017); Gropp 2017); for example, in biomedical research, only 20% of researchers perform 70–95% of the reviews (Klovanis 2016). (Tennant 2017)

These reviewing burdens are probably larger than people realize. We need to improve awareness. One solution might be to send periodic reminders to each person. The reminders could take the form of a monthly statement, like a bank statement, with an explicit accounting of recent (and long-term) activity. Activity includes both contributions (reviews) and deductions (accepted papers). If more people appreciated how much they are contributing and how much they are deducting, more people would be willing to do their fair share.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

2.7 Roles and responsibilities for middle management

So far, we have mostly discussed reviewing at the leaf of the tree. What about middle management (area chairs)?

1. Pushing papers down the tree
2. Popping the stack (as papers flow back up the tree).

2.7.1 Pushing papers down the tree

There are many ways to decide who reviews what. If a paper ends up with a random reviewer, then the paper is likely to be rejected because reviewers tend to abstain on papers that go beyond their expertise. As mentioned above, abstentions should not kill a paper, but they do.

As conferences become larger and larger, the tree becomes wider and deeper. More depth leads to more middle management (area chairs). Should middle management be involved in routing? That is, do we want to do the routing recursively or not? Should senior management (chairs) route recursively to subcommittees (area chairs) or directly to leaf reviewers? I tend to prefer recursion (to deal with scale) and to give middle management something to do.

There are many ways to route papers to subcommittees (or directly to leaf reviewers):

1. Delegate to authors (keywords)
2. Delegate to reviewers (bidding)
3. Delegate to middle management (manual assignments)
4. Delegate to software (automated routing; Yarowsky 1999)
5. Semi-automatic routing (automatic routing with manual post-editing).

There are conferences these days that do all of the above, with the possible exception of the last option (semi-automatic routing). Ironically, I am thinking the last option is the best option.

There are pros and cons to all of these suggestions. Keywords and bidding may not scale. Lots of authors do not understand what the keywords mean, especially when some of them are codewords for topics they have not thought about. Bidding is impractical when there are too many papers. It places too much burden on the reviewers. And worse, rumor has it that people can cheat with bidding by asking their friends to bid on their paper. This cheat is particularly unfair when the reviewing process is supposed to be double-blind.

Automatic routing scales well. I participated in the experiment reported in Yarowsky (1999). In fact, my manual assignments were the baseline for that experiment. (I used to take a couple of days to scan a couple hundred papers and assign them to a couple dozen committees by hand.) The experiment compared my performance to a text classifier that routed submissions to subcommittees based on the publications of the members of the various subcommittees.

IMHO, one of the failure modes of automatic routing involved conflicts of interests. The automatic routing had a tendency to route papers to the committee with the most expertise in the relevant area, and therefore, if there were any possible conflicts of interest, the automatic routing would send the paper there. I tended to think of pairs of committees. In addition to expertise, I was also on the lookout for conflicts. If there were too many conflicts with the best committee, I would send the paper to the second best committee. The automatic routing loved conflicts and never sent a paper to the second best committee.

2.7.2 Popping the stack (as papers flow up tree)

Middle management is responsible for mistakes (offensive reviews). They need to be on the lookout for reviews that would embarrass the organization if they showed up on social media. Middle management should also encourage reviewers to be as constructive as possible. If a reviewer does not like a paper, encourage the reviewer to teach the authors how to do better next time and avoid the temptation to say something unhelpful.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

Middle management should review the reviewers. Some organizations reward reviewers for doing a good job. Rewards come in many forms including bragging rights and $$. NeurIPS gives a few reviewers the rights to register for the conference (without having to win the lottery). Middle management is often involved in deciding who should receive such rewards.

Thank-you notes are an easy reward. Middle management should thank all the reviewers at the end of the process for their hard work by sharing with them the final results (along with the other reviews and meta-reviews). These thank-you notes are a teaching opportunity to help reviewers do better next time.

Middle management should encourage productive discussion. Too often, I receive mindless emails saying that my grade is different from some other grade, but there is little worth discussing because the paper is an obvious reject. When I do not like a paper, I might kill it with a grade of 5 (out of 10), whereas some other reviewer might kill it with a 1 (out of 10). Our scores might be different, but we are both trying to say the same thing. I am often more generous than other reviewers, but I know that a 5 is just as deadly as a 1 (though hopefully not as painful). In short, middle management should triage the discussion. We do not need to discuss easy rejects or easy accepts (unless we are nominating a paper for an award). Realistically, we cannot afford to discuss everything. We need to focus our limited resources sensibly.

It is (too) easy to average grades, but middle management should go beyond that. Averaging grades will favor boring (incremental) papers over interesting papers. A boring paper with three 7s (out of 10) will beat an interesting paper with an advocate, a second and an abstention. IMHO, we should reject a paper that does not have an advocate. The minimum bar should be an advocate and a second (and no serious objections). Boring papers tend to have lots of seconds but no advocates. We do not want boring papers. It is not good enough if the reviewers fail to find a flaw. Reviewers should be looking for something that excites them, and if no one can find that, that is a problem.

What do I mean by serious objections? In general, negative votes and abstentions should not count. In rare cases, the negative vote finds a serious flaw (like plagiarism) and convinces the advocate to change his mind. In those rare cases, I will take the negative vote seriously, but most negative votes should be counted like abstentions (no vote either way).

Why should we ignore negative votes? Too many papers are not even wrong. It is ok for a paper to wrong, but please do not be boring. A conference paper is different from a journal paper. Journal papers are archival. They should stand up to the test of time. A conference paper should make the audience think. If a paper gets the audience to throw rotten tomatoes at the paper, that is great (for the conference and the field, but maybe not for the receiving end of the rotten tomatoes). It is ok for a paper to be wrong as long as the exercise helps the community make progress.

3. What is the purpose of reviewing?

Reviewing needs to balance various competing and complementary pressures:

1. Audience: Make the program as strong (and interesting and credible) as possible. Remove iffy work (and especially cheating).
2. Authors: Reward success (publish or perish). Who is who in the field? Who deserves promotion? Authors want a process that is timely, credible (and not too onerous).
3. Scientific Society: Find ways to grow the membership over the long-term. Encourage new blood and new topics, especially in growth markets (with lots of students). Students are our future.
4. Reviewers: Make it as fun as possible. Reviewing should not be a thankless chore, but if so, we should do what we can do to manage the load (and reduce free-loading).
5. Industrial Sponsors and Funding Agencies: Publicity, recruiting, receive delivery on investments.

Can we come up with metrics to measure each of these different perspectives? The audience wants interesting work that makes them think. New ideas are more interesting than boring papers that make small incremental improvements over a boring paper we published at the last meeting. It is ok for a paper to be wrong (interesting new ideas often do not work out), but please do not be boring. Too many papers are not even wrong.

How do we measure "interesting?" I suggested in Church (2005) that reviewing ought to be a leading indicator of future citations. This is not always the case. The seminal paper on Page Rank (Page 1999) was rejected by SIGIR, even though it has received more citations than most of the papers they accept. Several leaders of the field pushed back (personal communication) and argued that that paper failed to meet their standards on evaluation.

I am not convinced that reviewers know better than the audience what is good for them. As the cliche goes, the customer (audience) is always right, especially when they are wrong.

Reviewing is very expensive and may not be worth the cost (Webber 2007). Should we seek ways to make it less onerous? Can we get it done quicker (List 2017)?

Should we bother with reviewing? If reviewing becomes too random, too slow, and too onerous, authors will find alternative solutions. There is already a tendency for authors to post papers on arXiv. As a practical matter, citations show up very quickly on Google Scholar. These days, Google Scholar might be faster than reviewing, and there may no longer be a need for reviewing as a leading indicator of future citations.

Google recently changed their ranking of venues by h-index, but under the old rules, arXiv was ranked higher than most other venues in physics, math (and computational linguistics):

> [T]he popularity and success of preprints is testified by their citation records, with four of the top five venues in physics and maths being arXiv sub-sections (Tennant 2017)

3.1 Purpose of reviewing: Catch cheating

Reviewing is intended to catch cheating and violations of ethics. Of course, there are lots of opportunities to cheat, not only for authors but also for many other parties in the system: authors can plagiarize, publishers can spam, and others can do other bad things. Some of the possibilities can be entertaining such as this[n] and this[o]:

Plagiarize,
Let no one else's work evade your eyes,
Remember why the good Lord made your eyes,
So don't shade your eyes,
But plagiarize, plagiarize, plagiarize. . .
Only be sure always to call it please, "research."

It is harder to joke about ethics.[p]

In any case, there is remarkably little cheating. Cheating is probably not that much of a problem (though this author and the editor of this journal have caught a few cases). In any case, there are other deterrents that are more effective than reviewing (Fox 1994). If we did away with reviewing, would there be more cheating? Do venues like arXiv have more cheating than venues with reviewing?

3.2 Purpose of reviewing: Promotion

Reviewing is definitely important for promotion decisions. Part of the reason for the inflation mentioned in Section 1.1 is grade inflation. Students need a huge number of publications to get their first job. To get your first job today, you probably need a publication record that used to be adequate for tenure. I was at an NSF workshop a few years ago where the senior researchers were encouraging people to publish fewer papers and spend more time on each paper. Students pushed back, arguing that they cannot afford to follow that advice. They believe, probably correctly, that

[n] https://www.vox.com/2014/11/21/7259207/scientific-paper-scam.
[o] https://www.lyrics.com/sublyric/5746/Tom+Lehrer/Lobachevsky.
[p] https://www.nytimes.com/2019/12/30/business/china-scientist-genetic-baby-prison.html.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

the old cliche, publish or perish, should be updated to publish every month or perish. This rat race is not good for the people stuck in the race (or anyone else), but it is what it is.

Should reviewing be as important as it is for promotion decisions? I am no longer impressed by a paper merely because it appeared in a highly rated venue. Citations are more impressive than venues. Papers that are highly cited but not published (like the seminal paper on Page Rank Page 1999) are more worth reading than papers that are published but not cited.

In the court of public opinion, what is most important for establishing a reputation? Currently, we cite papers in journals and conferences (and even arXiv), but we do not cite keynote talks, videos, podcasts, blogs, code on github, leaderboards, etc. At some point, success will depend more on content and less on format/venue, though admittedly, too many promotion decisions currently depend too much on format/venue, especially in certain parts of the world. Many of my colleagues in China, for example, feel enormous pressure to publish in top venues, leading to huge increases in submissions. The world would be better off if we could find a way to encourage people to spend more time on content, and publish less. But in the mean time, we are stuck with the cliche: whatever you measure, you get.

4. Conclusions

Reviewing is broken. We are well past denial. I no longer have to explain how broken it is. You all have your own war stories.

Negative feedback loops are effective. The reviewing process will improve over time if reviewers teach authors how to write better submissions, and authors teach reviewers how to write more constructive reviews. Please help improve the process by posting your war stories on social media (as I have).

We need to hold the officers on deck responsible for whatever happens on the ship. It is bad if they know about it, and worse if they do not. Program chairs (and area chairs) need to know that if not-ok reviews go out the door on their watch, their mistakes will show up on social media.

Given increases in submissions, scale is super-important. It has been said that scale makes it impossible to do a good job, but I am not convinced. We cannot afford to do reviewing badly (or else there will be even more submissions). We cannot afford to use unqualified reviewers (students).

The law of large numbers should work in our favor. Larger conferences should make it easier to find more qualified reviewers. One suggestion is to do a better job with accounting. How much is a fair share? Answer: 15 reviews per publication (not counting students, transients, and free-loaders). Statements of recent activity (like a bank statement) will help reviewers appreciate how much they ought to be contributing, and how much they have actually contributed.

There has always been a long tradition of bad reviewing (though we did not used to talk about it so much). We started SIGDAT/EMNLP in the 1990s to address problems with ACL reviewing. Now that EMNLP reviewing is no longer any different from ACL reviewing, it may be time to create a new venue with new reviewing practices.

It is possible that arXiv has already done this. The best work is showing up on arXiv, and citations in Google Scholar make it clear what is worth reading and what is not. Citations show up on Google in days/weeks, much faster than traditional reviewing (months/years). This alternative is not only faster than traditional reviewing but probably more fair and more effective as well.

Top universities know who is who in the field. If everyone cares about highly influential papers, talks, videos, blogs, github, and leaderboards, then that is what they will count. Unfortunately, there are lots of places around the world where bean counters on promotion committees have too much faith in precedent. They are still counting what they used to count (papers that no one reads). Given the reality that scale trumps precedent, they will eventually be forced to find a better way to make promotion decisions than counting minimal publishable units[q].

[q] https://www.chronicle.com/article/In-Defense-of-the-Least/44761.

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press

References

Chauvin A., Ravaud P., Baron G., Barnes C. and Boutron I. (2015). The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors. *BMC Medicine* **13**(1), 158.

Church K. (2005). Reviewing the reviewers. *Computational Linguistics* **31**(4), 575–578.

Church K. (2017). Emerging trends: Inflation. *Natural Language Engineering* **23**(5), 807–812, Cambridge University Press.

Fox C., Arianne A. and Vines T. (2017). Recruitment of reviewers is becoming harder at some journals: A test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review* **2**(1), 3.

Fox M.F. (1994). Scientific misconduct and editorial and peer review processes. *The Journal of Higher Education* **65**(3), 298–309.

Gropp R., Glisson S., Gallo S. and Thompson L. (2017). Peer review: A system under stress. *BioScience* **67**(5), 407–410.

Klovanis M., Porcher R., Ravaud P. and Trinquart L. (2016) The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PloS One*, e0166387.

List B. (2017). Crowd-based peer review can be good and fast. *Nature News* **546**(7656), 9.

Marcus M., Santorini B. and Marcinkiewicz M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**(2), 313–330.

Moher D., Galipeau J., Alam S. et al. (2017). Core competencies for scientific editors of biomedical journals: Consensus statement. *BMC Medicine* **15**(1), 1741–7015.

Page L., Brin S., Motwani R. and Winograd T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA.

Peters M., Mark M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In *NAACL*, pp. 2227–2237.

Smith R. (2010). Classical peer review: An empty gun. *Breast Cancer Research* **12**(4), S13.

Stelmakh I., Shah N. and Singh A. (2019). On testing for biases in peer review. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 5287–5297.

Tennant J., Dugan J. et al. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research, 6, Faculty of 1000 Ltd.

Tomkins A., Zhang M. and Heavlin W. (2017) Reviewer bias in single-versus double-blind peer review. *National Academy of Sciences* **114**(48), 12708–12713.

Van de Walle, F. (1986). *Infant Mortality and the European Demographic Transition*. Princeton, New Jersey: Princeton University Press.

Webber B. (2007). Breaking news: Changing attitudes and practices. *Computational Linguistics* **33**(4), 607–611, MIT Press, Cambridge, MA.

Yarowsky D. and Radu F. (1999). Taking the load off the conference chairs-towards a digital paper-routing assistant. In *EMNLP*.

Cite this article: Church KW (2020). Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering* **26**, 245–257. https://doi.org/10.1017/S1351324920000030

https://doi.org/10.1017/S1351324920000030 Published online by Cambridge University Press