Author’s response to reviews

Title: Quantifying Professionalism in Peer Review

Authors:

Travis Gerwing (t.g.gerwing@gmail.com)
Alyssa Gerwing (am.gerwing@outlook.com)
Stephanie Avery-Gomm (stephanie.averygomm@gmail.com)
Chi-Yeung Choi (c.choi@uq.edu.au)
Jeff Clements (jefferyclements@gmail.com)
Joshua Rash (jarash@mun.ca)

Version: 1 Date: 25 May 2020

Author’s response to reviews:

We would like to thank the editor and two reviewers for their thoughtful and insightful comments. We have used feedback to substantially improve the quality of the manuscript. We have included detailed responses to reviewer concerns below.

Editor’s Comments

Comment 1: “likely impact of such comments on the mental health of academics – I prefer the term you used in the introduction - psychological distress – and would advise you use that one here too”
Response: The term “psychological distress” has been adopted in the revised manuscript (Lines 33-35).

Comment 2: “please provide an overall average for indicators as you did for unprofessional comments”
Response: This value (41%) was added, Line 27.

Comment 3: “ please add in the methods of the abstract that all reviewer sets were from fields Ecology and Evolution and Behavioural Medicine”
Response: This text was added, Lines: 23-24.

Comment 4: “ please provide scientific fields and countries of the included 6 early career researchers.”
Response: We added the scientific field of case studies, however, to protect identities we have not identified countries (Lines: 77-82).
Comment 5: “please describe how did you find the 327 emails - with full details of sampling and if any sample size calculations were done”
Response: The manuscript has been revised to include a clearer description of sampling procedures, Lines: 77-82:
No sample size calculations were performed given the descriptive nature of the study. Our goal was to obtain as many manuscripts and reviewer comment sets as possible, but ideally ~300 manuscripts.

Comment 6: “please specify which program did you use for randomization, what was the total number of reviewer sets in those categories, and why this specific number was chosen
Response: Random numbers were generated in Microsoft excel, and this information was added to Line: 90.
Unfortunately, we can not report the total number of reviewer comment sets available on Publons given that this number changed daily as more reviewers were added. However, search results were saved, allowing resumption of evaluation without assignment of random numbers being impacted.
We have clarified the number of manuscripts we attempted to review, Lines: 89-92.

Comment 7: “ If I am not mistaken, there is no need to call these reviewer sets – you could use just say reviews.”
Response: We would prefer to retain “comment sets.” Early versions of this manuscript used the phrase “reviews,” but this created confusion regarding the nature of reviews, and how multiple reviews were counted. While slightly awkward, this phrase enhances clarity. We will adopt the terminology “reviews” if the editor insists.

Comment 8: “Additionally, did you indicate in your database if the reviews were first round or second/third – it could be interesting to see if more bad comments were received in initial or revised rounds.”
Response: We agree that this analysis would be interesting. While our database does contain this information for some reviews, we would prefer not to conduct this analysis. Not all reviewer comment sets were uploaded to Publons (already mentioned on lines 267-270). When evaluating reviewer comment sets, mention to other rounds of revisions that were not available occurred frequently enough that we were not comfortable performing such an analysis. As we detail in our manuscript (Lines: 267-269) this limits our ability to discuss responses to specific papers, but when reviewer comment sets are pooled, it enables us to speak in generalities about the experience of peer review. We could perform such an analysis on the case studies, but as the sample size is limited, we feel that adding an extra component to this paper would decrease clarity and distract from the overall results.

Comment 9: “Dot missing at the end of sentence - and 1491 reviewer comment sets”
Response: The period has been added, Line: 92.

Comment 10: “ while for reading I understand the reason for blinding, for reporting it would be beneficial if you could state how many different journals your whole sample covered, and if you could check for any time trends, additionally, as you did not check reviewer names – it could be that several bad apples are behind many of those comments”
Response: Unfortunately, to ensure that our opinion of journals did not influence how we viewed reviews, we did not record journal names nor years of publication. Further, we wanted to ensure that no details in our manuscript could be used to identify reviewers, or match case studies to authors. While such an analysis would add greatly to this conversation, we felt that preserving reviewer anonymity was more important.

Comment 11: “As you divide the results into unprofessional comments and indicators of superficial reviews, please use the same structure, and numbering here, 1 for unp... 6 for indicators (1 to 6)”
Response: This change was made (Lines: 120-176).

Comment 12: “this young lady is lucky to have been mentored by the leading men in the field” – without knowing the rest of this comment, not sure why on its own this would be unprofessional or bad – it could have been a praise and a compliment for a nicely written paper – please provide more details about this comment”
Response: There is a subjective element to determining whether a comment is demeaning. This comment was provided with very little context in the review, but the subsequent two demeaning comments in the supplemental table were also extracted from that review. This comment was identified as demeaning by the authors given that the gender of a scientist is not an appropriate subject for comment. This raises the issue of misgendering a scientist, dead naming them (using an incorrect name following transition), etc. As many specific examples of demeaning comments could be included or not depending on personal perspectives, we have elected to retain this comment as it was unanimously identified as insulting by female scientists surveyed.

Comment 13: “as for unprofessional category above, I feel actual examples with quotation marks would better serve the purpose and follow the previous paragraph style”
Response: This change was made (Lines: 122-127)

Comment 14: “How were you able to asses that the stamen was actually stated? Due to rebuttal letters or? I advise using an example with both the statement and the part with the proof it was stated. And please provide a rationale why was this only assessed in case studies?”
Response: Yes, this was evaluated by assessing author responses and when possible comparing to manuscripts. We only assessed the validity of comments in case studies, as not all Publons reviews included a response by the authors. We have clarified this on Lines 138-144 of the revised manuscript.

Comment 15: “This is to me a very tricky aspect of your analysis – Did the commenters say why it was wrong without providing references? Or considered them to be such general knowledge that no citation was needed – I would highly welcome more detailed examples of these. Also consider the comment I also made below for missing literature, it applies here too.”
Response: As described between Lines: 146-154, we included examples where a reviewer raised concerns, without providing references or sufficient text to support this point. We now provide examples of this in the manuscript, Lines: 146-154.
We deem these statements unhelpful given that they offer no guidance for the authors, nor do they allow for an evaluation of their validity. As professionals, it is inappropriate to simply say that something is wrong, without defending that assertion. Rather, references could be provided
to guide the authors. For instance, if an ANOVA was done incorrectly, it is relatively easy to point out what was done wrong, instead of only stating that the analysis was poor. The comments we identify here do not provide enough detail for the reviewer’s claim to be assessed. We have expanded on this point in the revised manuscript (Lines: 146-154).”

Comment 16: “of them to provide links or references, but as you are here denoting the practice – you have not checked how many times reviewers did provide the links.” Response: The first part of this comment appeared to be missing. If this comment is referring to editor’s removing reviewer self citations, we are confident this did not occur frequently in the comments identified here. Comments we included were statements such as: “No contrasts with other studies! They are either behind latest advances or left these studies out to inflate their own impact,” and “literature was inadequate, expand.”

Comment 17: “This one also troubles me – who is unaware here – the reviewer – if so, is this not rather an indicator of an editor choosing a reviewer who isn’t the right person for this article. Also, even some commonly used methods still have very known problems, ad if a reviewer is pointing this out – why is this bad?” Response: The editor is correct that these issues could be a result of an editor selecting a reviewer who is unqualified to comment upon certain aspects of the manuscript. However, this does not change the fact that the comment itself is of a troubling nature. We now clarify this point (Lines 156-161). Comments identified as contradicting well supported methods did not include instances where reviewers asked for nuanced justification of methods (e.g. I am wondering why “x” was used instead of “y”). Rather, we counted instances where reviewers viewed common methods as strikes against the manuscript. For example, arguing that the use of a 250 µm sieve to study benthic invertebrates should result in summary rejection, despite over 50 peer-reviewed citations from the past 30 years supporting its use.

Comment 18: “Another tricky one, the whole point of the criticism might have been that the researchers did not do a good literature search, and therefore the reviewer does not intentionally want to help them, but want them to do the work themselves. You can debate that it would be nicer if they did, but we are not sure are they asked to do so by the journal, and what is their stance on this.” Response: We expect this point to be one of our most controversial peer review issues, and many academics may disagree with us. Hopefully, this will spark a conversation about what makes peer review most beneficial. However, we maintain that if the reviewer is indeed an expert in the subject and critical literature is missing, it would be easy for them to provide even a single citation. We agree that while it would be nice for reviewers to do this, we do have an evidentiary concern with such comments as well. When such claims are made without providing even a single citation, the reviewer’s claim that literature is missing can not be assessed by the editor. No evidence is provided to support such a claim, and in many responses, we saw authors successfully argue that literature was not missed. We now clarify this between Lines: 163-167.

Comment 19: “1) can you for this one especially, but also for all the categories, provide the average number of words per review – should be easy to calculate. While I understand that a short review might seem superficial – if as a reviewer I detected a huge flaw in the paper – I
would not bother with the rest – so I would advise you check this category and split it in two - those that mention a serious problem, and those that do not.”

Response: We were careful not to identify short reviews as superficial if they provided detail. For instance, if a fatal flaw was detected and described in a short paragraph, or even a few sentences, this review was not coded as superficial. However, if fatal flaws are vaguely referred to but not detailed, then those reviews were coded as superficial. We now clarify this between Lines: 169-176.

Comment 20: “this young lady is lucky to have been mentored by the leading men in the field – u used this one twice, I would advise against using the same examples, also see my above comment for this particular comment”
Response: Tragically, we have many examples to draw upon. We have replaced this example (Lines 185-186).

Comment 20: “I would rather you did not provide ranges as you do now, but show only the averages, and then in the paragraph after say, that you found differences between fields, and point to supp. Table 2.”
Response: This change was made (Lines 178-194).

Comment 21: “Do not use ~, say almost.;
Response: This change was made (Line 196).

Comment 22: “Additionally, I advise rephrasing the first two sentences into something like: After reading almost 1500 reviews, our study has found X percentage of unprofessional comments, and X with at least one indicator of superficial reviews for studies published in Ecology and Evolution and Behavioural Medicine. This, personally troubles us, as if put bluntly it means one can expect to receive….”
Response: We have edited the opening sentences in light of these suggestions (Lines: 196-200).

Comment 23: “I did not go into details with the rest of the discussion, as I am looking to hear back from you on the above comments I made. I will only two more notes; 1) that you should include a limitations section in your discussion, 2) please reflect on research conducted by Hyland and Jiang https://doi.org/10.1016/j.jeap.2020.100867”
Response: We have included a detailed limitations section within the revised manuscript (Lines 314-342).
Thank you for providing that reference, we have incorporated it throughout the manuscript.

Comment 24: “As the unit of analysis is a single review – how did you count if the same review had several of the categories – It would be very helpful if you would say the percentage of reviewers that did not have a single “bad” remark and those that had more than 1.”
Response: Each question was assessed independently for each comment set. A comment set could code positive for questions 1, 3, and 5, but not 2, 4, and 6. That was the basis for presenting our results by unprofessional comments and indicators of superficial reviews. We have clarified this in the revised manuscript (Lines 111-119).
Requested information was added Lines: 178-182.
Comment 25: “I believe that your Supplementary Table 1 better shows the results than Figures 1 and 2, and would advise they become supplements, while the table becomes the main table. Do correct the spelling of Publons in the table.”
Response: This change has been made.

Comment 26: “Finally, some of the examples you provide in the supplement are not offensive or demeaning in my opinion – all of those related to language: carefully proof-read by a native speaker This writing in the introduction is not high quality and should be reviewed by a native English speaker As English is clearly a problem I suggest other authors give this a good read Authors should ask for help of a colleague whose first language is English You might suggest authors how to praise this better, if you did find examples on the same topic that you did not classify as bad, but to me it is not offensive if you point out that the language needs editing, so all of these should be taken out, and perhaps you can create a new category about language editing.”
Response: With respect, we disagree strongly with the editor regarding the offensive nature of these comments. Such comments are inherently dismissive and suggest that scientists for whom English is not their first language produce writing that is intrinsically of lower quality than native English speakers. Such an assumption is unprofessional, and in many cases these comments incorrectly identify the native language of the author. There is a degree of subjectivity to coding a reviewer comment as unprofessional. This is exemplified by our disagreement with the editor on this point (many people will likely disagree with us). We have elected to include these comments as to us, and to those who received them, they are viewed as dismissive and hurtful. Even comments that are not intended as unprofessional can reliably be interpreted as such and have negative impacts upon the recipient. Such language can easily be avoided. Guidance for authors is included in the revised manuscript (Lines: 285-289).

Reviewer 1

Comment 1: “differences between external and internal peer reviewers: It is not clear which percentage of the studied peer review reports are written by internal peer reviewer such as editorial boards or editorial office members. The handling editor might have a more relaxed approach toward such groups of referees and not necessarily read the entire body of the reviewer comment.”
Response: We evaluated all comments provided to authors with the same rubric, as the source of a comment should not influence the professionalism nor technical merit of that review. Reviewer 1 is correct that differentiating between internal and external review content is an interesting follow up to this study. We have added the following text to the limitation section to help guide future research “Second, reviewer behaviour was pooled across external and internal peer-reviewers which prevents a nuanced understanding about whether reviewer behaviour varied among those on the editorial board.” (Lines 315-342)

Comment 2: “differences between the peer review policies of journals: It is not clear which peer review policies (single/double/tripple blind or open peer review) have been considered. This impacts a particular category about missing link to a required article. In case of double-blind
policy, it might be that the reviewer originally requested authors for a self-citation which breaches the peer review policy of the journal. Editors/editorial offices of double blind journals usually scan the content to remove reviewer’s self-citations.”

Response: We did not extract or evaluate journal names in order to preserve blinding and prevent expectancy effects. Unfortunately, this does not allow us to differentiate between various peer review systems (Lines: Lines 315-342). We have added the following text to the limitation section of the revised manuscript “Third, journal names were not extracted or evaluated in an attempt to preserve blinding and prevent expectancy effects which prevented a nuanced understanding about whether reviewer behaviour varied by journal, impact factor, or peer-review policy (e.g., single/double/triple blind or open peer-review).”

We are confident that censorship of self citations did not occur frequently in the comments identified as problematic. Examples we included were statements such as: “No contrasts with other studies! They are either behind latest advances or left these studies out to inflate their own impact,” and “literature was inadequate, expand.”

Comment 3: “subject area limitations: While it is understandable that authors chose subject areas closer to their field, the lack of study on other fields limits the presented results. To put this in perspective, it is important to mention how big the size of these specific communities are since peer review is a human interaction and people behave differently in closer/small communities than in wider groups.”

Response: We agree that expansion of our analysis into other fields would provide greater insight into the experience of peer review. We were forced to limit the focus of our analyses as there are an unending supply of reviewer comments to assess. However, by assessing 1500 reviewer comment sets, we believe that the identified trends will be informative for all subjects. We hope that our work inspires other authors to do similar analyses in their fields.

It was not possible to quantify the size of the scientific communities as suggested by Reviewer 1.

The following text has been added to the limitations section of the revised manuscript “Fourth, reviewer behaviour was only considered for “Ecology and Evolution” and “Behavioural Medicine.” Both are large fields and it was not possible to determine how close the scientific community was within areas of sub-specialization. Peer review is a human interaction and reviewer behaviour could vary by the size or interconnectedness of the community.”

Comment 4: “anonymity: Although it is stated that to respect anonymity authors did not control for reviewer/author gender, ethnicity, academic ranking, I don’t think running such detailed studies provided having a properly defined categories, would harm anonymity. It would be quite a complementary and insightful result to know if younger/women/global-south affiliated authors receive harsher reviews for example.”

Response: We have elected to avoid such analyses due to an abundance of caution to maintain reviewer anonymity, as well as to protect the identity of authors who provided case study reviews. Further, the analyses suggested by Reviewer 1 are impossible to perform with the available datasets. For all published papers, the only available information about authors are their names, institutions, and in some cases, degrees and positions. Gender, ethnicity, age, and sometimes academic rank would need to be inferred from a name which would be prone to error.

We have added the following text to the limitations section of the revised manuscript: “Fifth, neither the authors gender, ethnicity, nor academic rank were quantified and assessed in the
Comment 5: “Correlation with recommendation type, editor decision, and resubmissions: The study does not address which percentage of the referee reports are correlated with ultimate recommendation, i.e., accept, minor/major revision, reject. It would be insightful to see the correlation of 'unprofessional' reviews with negative recommendation types. It is also not clear if these authors have received their 'unprofessional peer review' for the first time submission or in follow up submissions after being rejected from a different journal.”

Response: Unfortunately, recommendation data was not available for the majority of Publons reviews. Similarly, not all reviewer comment sets were uploaded to Publons (mentioned on lines 265-270). When evaluating reviewer comment sets, mention of other rounds of revisions that were not available occurred frequently enough that we were not comfortable performing such an analysis. We have added the following text to the limitations section of our revised manuscript “Sixth, final recommendation (i.e., accept, minor/major revision, reject) was not available for many reviewer comment sets which precluded a nuanced understanding of the degree to which reviewer behaviour varied by final recommendation.”

Comment 6: “article types: it is not clear for which article types these collected peer reviews were submitted, i.e., researcher article, review article, case study, etc. This needs to be clearly mentioned.”

Response: We have elected not to present this information as we believe that article type should not have an influence upon reviewer professionalism nor the technical merit of the review. While each article type represents unique challenges to authors and reviewers, we do not see how this would influence the experience of peer review.

Comment 7: “non-exclusive attributions: none of the 7 listed categories are truly defined and considered as exclusive. In fact the first group, 'unprofessional' covers almost the rest of the group. There is another study very similar to what is presented here that is specifically classifying such attributions in a non-exclusive manner. It would be beneficial if authors check that study and try to replicate using the same method (reference: https://doi.org/10.1016/j.jeap.2020.100867)”

Response: We agree with Reviewer 1 that there is overlap between some questions; however, this is difficult to avoid. For instance, a comment could be unprofessional while at the same time being an argument from authority. However, we do not view this as a limitation of this work. Instead, each category provides a different perspective on peer-reviewer comments, illuminating different potential issues. Individually each of the questions only provides one perspective, but when taken together they offer a holistic view of potential peer review issues.

Thank you for providing that reference, we have incorporated it throughout our manuscript. However, the source of reviewer comments in Hyland and Jiang (1) is a website “shitmyreviewrsays” hosted by tumblr. This website provides comments that are short sections of a review, provided without context, and submitted by authors. While this paper provides great insight into peer review, it is difficult to compare our methods to theirs, as we examined 1500 entire reviewer comment sets, not 850 comments consisting of 1-2 sentences. Finally, the goals of the two papers are divergent. Our papers starts with 1500 reviewer comment sets that may or may not be problematic. Hyland and Jiang (1) start with 850 comment sections that have already
been identified as problematic and assess the comments themselves. Conversely, we quantify the incidence of unprofessional comments within a broad subsection of reviewer comments. As such, utilization of Hyland and Jiang (1)’s methods are not appropriate to our objective.

Comment 8: “inter-rater agreement: it is not clear if one or all authors read the entire body of the collected referee reports and mapped to these categories, or there were some inter-rating mechanism in place.”
Response: We have revised our submission to clarify that reviewer comment sets were coded by TGG and JR. A coding rubric was developed a-priori and pilot tested on 5% of reviewer comment sets to ensure agreement between coders. This coding rubric will be made available through reasonable requests made to the corresponding author. Inter rater agreement was substantial following pilot testing, and each of the remaining reviewer comment sets were coded by only one author. We recognize that electing not to code every reviewer comment set in duplicate is a potential weakness and have added the following text to the limitations section of the revised manuscript “Seventh, the majority of reviewer comment sets were coded by one investigator which may raise concerns about reliability of codes despite substantial inter-rater agreement during pilot testing the rubric across 5% of reviewer comment sets.”

Comment 9: “Comparison between two samples: It is not clear what the subject areas of the 6 PIs who voluntarily shared their history of peer reviews are. Also, it is not mentioned which proportion of shared peer reviews by this group is related to what recommendation term. As a result in conclusion we miss a clear reason why this group has received more 'unprofessional' referee report and which period of their career. Where these reports received from the same group of journals or they changed in the course of time?”
Response: Subject areas of the case studies was included in Supplemental Table S1, but now are included in Table 1 and on Line: 81.
We did not analyze peer review responses based on final recommendation, refer to our response to Reviewer#1, Comment#5.

Reviewer 2

Comment 1: “The issues addressed by the authors are extremely important in modern research, and vastly underappreciated. Abusive an inappropriate review is a huge and serious problem in academia, particularly with regard to ECIs. In broad strokes, a systematic review of inappropriate review comments is an excellent way of bringing these issues to light.”
Response: We would like to thank Reviewer 2 for the kind words. We feel as through it is pertinent to note that our manuscript represents a descriptive quantification of the prevalence of inappropriate reviewer behaviour during the peer-review process rather than a systematic review per-se. Unfortunately, there is insufficient evidence in the extant literature to perform a systematic review at this time. We have revised our stated study objectives to make this point clearer (Lines: 56-66).

Comment 2: “However, while the issue and broad idea is generally excellent, I have major concerns about the design, execution, and reporting. I was unable to follow the systematic process of the methods, nor was I sure about the justification of many of the key methodological
decisions. It appears that this was a qualitative or mixed quantitative / qualitatively coded study, but neither the qualitative nor the quantitative frameworks or methods were described, if any such frameworks were utilized. There also appears to be writing issues that made it extremely difficult to follow and understand what was done, and what the results were.”
Response: We have rewritten the methods and results sections extensively. Please refer to our detailed responses to the editor’s comment, as well as the more specific comments from Reviewer 2 below.

Comment 3: “Finally, I have ethical concerns about the conduct of the case studies, which may be in violation of human subjects research requirements.”
Response: We appreciate Reviewer 2’s concern about ethics with regards to this subject. To the best of our knowledge, there are no formal ethical requirements, nor approved protocols for reviewing peer-review comments. Numerous studies that discuss, and report peer-reviewer comments while maintaining anonymity have been conducted and published without need for institutional ethics approval [e.g. Beaumont (2), Silbiger and Stubler (3), and Hyland and Jiang (1)]. Still, we would like to err on the side of caution and are in the process of consulting with the head of research ethics at our institution about this matter. We did not wish to delay the resubmission process, and agree to act in accordance with advice that we receive (i.e., we agree to obtain appropriate ethical approvals during the next round of reviews if we are advised of this course of action).

Comment 4: “To review this section, the criteria I am using is that I - an experienced but otherwise naive researcher - should be able to understand well enough to at least roughly replicate what the authors did. Unfortunately, after several re-readings, I was unable understand the process of how the data were acquired and organized. This is in part due to inconsistent writing and language, but also in part due to incomplete information. Using headings and subheadings, and text that explicitly said that these were independent searches would go a long way, as would diagrams explaining the process.”
Response: We have rewritten the methods section extensively to be clearer, please see our responses to the editor’s feedback above.

Comment 5: “I would strongly suggest flow charts or similar. In the results section, diagram types that might be found in systematic reviews, such as PRISMA diagrams, would go a long way to help guide the reader through the process.”
Response: We have included a flow chart in the revised manuscript that depicts the inclusion of comments sets across case studies/Publons, and area of study.

Comment 6: “As best I can tell (but I am still honestly not sure that this is the case), there were two completely independent sources of review data: "author case studies" and Publons reviews. If that is the case, this should be spelled out explicitly. A full paragraph or similar laying out the data collection process would help greatly.”
Response: This information has been added to the revised manuscript (Lines 68-74).

Comment 7: “Author case studies: The case studies seem out of place, and I am unsure they add to the study. I am concerned that this will produce a highly selected group of individuals who experienced particularly brutal reviews (people who have worse reviewers will be more likely to
volunteer, heavily biasing the sample). As far as I am aware, this is not discussed in either the methods or discussion sections. In general, I am not sure why these were included at all, given that you have a systematically selected sample elsewhere. One suggestion for this section would be to clearly lay out what you are trying to accomplish with the early career author case studies, and structure the results around these justifications.”

Response: We feel that it is imperative to maintain the author case-studies as well as reviews retrieved through Publons. It is true that early career investigators who elected to participate may represent a biased sample of individuals who have received a preponderance of unprofessional reviews. It is also true that reviews published on Publons may represent a biased sample of reviewers who provide particularly courteous reviews (i.e., those who agree to make their reviews open access may censor comments more carefully). The true experience of peer review is likely somewhere in the middle, and including both groups helps us better elucidate this trend. We have added the following text to the limitations section of the revised manuscript: “Finally, results may be suspect to selection bias. Authors who agreed to participate in case studies are likely those who are most passionate about the issue of peer-review and may have been more likely to experience unprofessional reviewer comments, while reviewers who agree to make their reviews public on Publons may be more courteous in their responses than those who do not.”

Comment 8: “How were the original 327 researchers selected to be e-mailed? What was in the e-mail that was sent to them?”
Response: We have expanded our description of the email selection process (Lines 77-82). Each email was written individually, and briefly laid out the goals of the project, asking researchers if they would be willing to contribute their peer-review comment sets to the process.

Comment 9: “I am also concerned about ethical consent, which is not at all addressed. Was consent obtained from these individuals? What about the reviewers of the articles whose reviews (abysmal though they may be) were written under the assumption that they would not be distributed. These are non-public data of a sensitive nature that are being used for research purposes in a potentially vulnerable population, which at minimum necessitates specific justification for why it would be exempt from human subjects review.”
Response: Refer to our response to Reviewer 2 Comment 3. It is important to note that Publons is an open access online repository of peer review comments. As such, no ethical constraints exist regarding our use of this data, nor can there be any expectation of anonymity for reviewers identified in this database. We exceed this basic threshold by ensuring that anonymity of reviewers from Publons is protected in our study.

We are also unsure whether peer-reviews are written with the expectation that they would not be distributed. Review comments are expected to be seen by the editor, handling editors, and editor in chief. Further, the peer-review comments are emailed to authors. Finally, many journals then openly email reviewer comments to peer-reviewers once a decision has been made. Given the open communication of reviewer comments already occurring, to say nothing of researchers then passing around comment sets between themselves, we maintain that there is no realistic assumption to not distribute reviewer comments. However, we do believe that there is a realistic assumption for anonymity, as most peer-review is designed to be anonymous. That is why we adopted measures to protect reviewer identity, even if this limited the analyses we could conduct.
Comment 10: “Publons: Why were these three journal types selected? This seems simultaneously like an arbitrary but also very specific set of sections. If you were looking for diversity, why not a random selection from all of Publons' subject areas?”
Response: These subjects were chosen as they corresponded to the case studies, thus enabling a more direct and informative comparison.

Comment 11: “There are also quite a few results in this section. The methods should contain the search strategy, while the results should contain the counts of articles contained at each stage of the process (preferably with a PRISMA-like diagram).”
Response: We have rewritten the methods section extensively to include this information. We have included a flow diagram in our revised submission.

Comment 12: “Reviewer comment sets: The definition for this appears after it is used several times, which is very hard to follow. Please integrate this into the text. You also describe an "interaction" but do not define what this means. It would help a lot if you gave a basic description of what Publons is, and the data therein.”
Response: This change was made (Lines 68-74).

Comment 13: “This is the first mention of "coding" in the paper, which is typically a qualitative concept and framework. However, no mention is made on the qualitative methods for how the codes were determined. What qualitative frameworks were used, if any? I am concerned that the lack of mention here suggests that no such frameworks were used or considered, or that a qualitative research expert contributed to the methods design and execution.”
Response: This information is detailed in the revised manuscript (Lines: 111-119).

Comment 14: “Did TGG and JAR have an agreed-upon coding language beforehand? Was there any cross-validation of the coding?”
Response: Refer to our response to Reviewer 1, Comment 8.

Comment 15: “These appear to be categories which are extremely subjective, and some appear to take quite a bit of effort and knowledge from the author coders. For example, what is an "unprofessional" comment vs an acceptably professional comment? I am not aware of any standard definition here, and this is the key argument of the paper. How was this determination actually made? Subjectivity is necessary, and even desirable here, but I have no confidence that I would be able to replicate this coding were I to attempt it.”
Response: We have expanded this section greatly to provide more details and enhance reproducibility (Lines: 121-176).

Comment 16: “I see no information on how the statistics were generated. Elsewhere in the paper, there are ranges of proportions which I assume (I did not find a label) are 95% confidence intervals generated somehow, but how?”
Response: This information is now detailed between Lines: 111-119.

Comment 17: “The results are extremely minimal and lacking in detail, and missing critical information. For starters, we need some basic descriptive statistics here about the sample. How
many total comments were coded, and from what sources? The total number appears in the Discussion, but needs to be in the results.”
Response: We have expanded the results section and provided greater detail (Lines: 178-194).
Descriptive statistics of the sampled reviewer comment sets were already provided between Lines: 101-102 in the methods, as well as in the caption for Table 1.

Comment 18: “It also appears as though the authors pooled the case studies and the Publons reviews together. These should be treated separated (or alternatively the “case studies” removed from the paper altogether), unless there is a good justification for it.”
Response: As shown in Table 1 (supplemental S2 in the previous version) we present pooled results as well as divided by subject, and case study/Publons.

Comment 19: “I am very confused about what (I assume) are the 95% confidence intervals here. These appear to be extremely wide for a sample size of 1500. However, I have no way of verifying that this is a problem, as I am not sure what these numbers represent. What does this range represent?”
Response: This was a point where we could have been clearer. Values indicated mean estimates and ranges (min, max). This text has been removed on the editor’s suggestion.

Comment 20: “I would strongly suggest a paragraph for each of your seven sections, including quotes. I suspect that the quotes themselves are likely the most valuable part of this paper, rather than the counts, and it would be great to see them. Much of these data exist, but in the discussion section and not in the results section.”
Response: Quotes are presented in Supplemental Table S1. We have also expanded our methods section to include examples for all questions.

Comment 21: “I would strongly suggest modelling this section after common structures in the qualitative/mixed methods literature.”
Response: We are unclear what Reviewer 2 is asking us to model. The purpose of this paper was to describe the incidence of common peer review issues in a large subset of reviewer comment sets. To protect anonymity of reviewers, and to conduct this research ethically, we did not collect any information that could identify authors or reviewers. Not only do we lack the variables needed to create such a model, but creation of such a model is beyond the scope of this paper.

Comment 22: “A huge amount of the discussion section are results that should be in the results section. For example, the quotes, the statistics, etc., are much better suited to the results section, as these are the product of the methods. This partially explains why the results section is lacking.”
Response: The results section has been expanded greatly to include this information, and results have been removed from the discussion section.

Comment 23: “Much of the discussion section appears to make broad and wide sweeping arguments regarding the publication system. While I believe they are well-intended, the paper would benefit from a much narrower focus. For example, the authors describe "post truth" as a relevant concept, which is tangential to the topic at hand, adjacent at best. I was unable to follow the line of logic that lead us from abusive comments to post-truth society at large. The authors
are, in my opinion, attempting to do vastly too much here, and not all of it relevant to the actual study performed. For reference, there are 6 pages of discussion section for one paragraph of results. To me, this alone suggests that the authors should reconsider their broader argument, and keep the discussion relatively narrowly to the paper at hand.”
Response: We have removed this text, and rewritten the discussion/conclusion section to align more closely with the main thesis of the manuscript (Lines: 344-365).

Comment 24: “Ironically, I found much the language in this section to, itself, be potentially unprofessional. For example, much of the language is heavily tinged with moral arguments and the authors' beliefs. While I don't disagree with any of them off hand, opinions and preferences should be clearly marked with "we believe," "we suggest" or similar. For example, there are large sections describing what reviewers "should" do. In general, I agree, but it is best to state these as the beliefs of the authors.”
Response: We have reviewed our language throughout the manuscript and been careful to clarify our opinions and beliefs as such. We have also removed text from the discussion that was more heavily tied to moral arguments.

Comment 25: “Similarly, I found the dismissal of key limitations as "obvious criticisms" rather than key limitations to be poorly worded. It may not be "obvious" to many, nor are they necessarily criticisms, but rather limitations.”
Response: We have included a detailed limitations section to the revised manuscript. Refer to our response to comment 23 from the Editor. We have also rephrased this sentence (Line 278) to remove the word obvious.

Comment 26: “Many key limitations were missing from this section, including the selection bias induced by self selection of the "case studies," lack of comparison with non-ECR-generated papers, subjectivity of the assessment of the questions, ethical concerns, etc. A more thorough assessment of the limitations of this paper would help greatly. I was also not fully able to assess these limitations, in part because I didn't fully understand the methods.”
Response: We have included a detailed limitations section to the revised manuscript. Refer to our response to comment 23 from the Editor. Finally, we address the Reviewer’s ethical concerns in Comment 9.

Literature Cited in this Response

