Leading Edge
Commentary

Cell

Universal Principled Review: A Community-Driven Method to Improve Peer Review

Matthew Krummel,[1,*] Catherine Blish,[2] Michael Kuhns,[3] Ken Cadwell,[4] Andrew Oberst,[5] Ananda Goldrath,[6] K. Mark Ansel,[7] Hongbo Chi,[8] Ryan O'Connell,[9] E. John Wherry,[10] Marion Pepper,[5] and The Future Immunology Consortium

[1]Department of Pathology, ImmunoX Initiative, and Parker Institute for Cancer Immunotherapy, University of California, San Francisco, San Francisco, CA 94143, USA
[2]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Department of Immunobiology, The University of Arizona College of Medicine, Tucson, AZ 85724, USA
[4]Kimmel Center for Biology and Medicine at the Skirball Institute and Department of Microbiology, New York University School of Medicine, New York, NY 10016, USA
[5]Department of Immunology, University of Washington, Seattle, WA 98109, USA
[6]Division of Biological Sciences, Section of Molecular Biology, University of California San Diego, La Jolla, CA 92037, USA
[7]Sandler Asthma Basic Research Center and Department of Microbiology & Immunology, and ImmunoX Initiative, University of California San Francisco, San Francisco, CA 94143, USA
[8]Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[9]Huntsman Cancer Institute and the Division of Microbiology and Immunology, Department of Pathology at the University of Utah, 15 N. Medical Dr. East, Salt Lake City, UT, 84112, USA
[10]Department of Systems Pharmacology and Translational Therapeutics and Institute for Immunology, Perelman School of Medicine, and Parker Institute for Cancer Immunotherapy, University of Pennsylvania, Philadelphia, PA 19104, USA
*Correspondence: matthew.krummel@ucsf.edu
https://doi.org/10.1016/j.cell.2019.11.029

Despite being a staple of our science, the process of pre-publication peer review has few agreed-upon standards defining its goals or ideal execution. As a community of reviewers and authors, we assembled an evaluation format and associated specific standards for the process as we think it should be practiced. We propose that we apply, debate, and ultimately extend these to improve the transparency of our criticism and the speed with which quality data and ideas become public.

Introduction

The peer-review process can be an inefficient part of the scientific endeavor. A recent study estimates that 15 million hours of effort are squandered each year in the process of peer review (Rubriq, 2013). This is most likely a dramatic underestimate. A large portion of this delay may be attributed to peer reviewers' insistence on proposing "reviewer experiments," which take time and effort to complete and often fail to fundamentally alter a manuscript's conclusions (Ploegh, 2011). Delays are also magnified by pressures for authors to draw profound conclusions from data, which will often fall short of "proof beyond the shadow of a doubt" (Kaelin, 2017). The inability to then unambiguously support a "game-changing" conclusion and extend it to all scenarios leads reviewers to feel the need to require new experiments in an attempt to warrant such grandiose conclusions.

The primary roles of the peer-review process should be to vet the quality of the data using field-specific criteria and to request a balanced discussion of its validity and meaning. Peer review is also important insofar as we have "tiers" of journals. Here, it is ultimately used for gatekeeping, and reviews become more subjective, focusing upon the extensibility of data and its possible impact. In the worst cases, peer review takes the form of a diabolical game in a contest of ideas—a fruitless diversion in the long term and one that hinders the airing of data and/or opinions that disagree or appear to disagree with previous ideas. Frequently, we individually feel that comments of a certain nature (e.g., the legendary "reviewer #3") should not be permitted, but we as a community have not adequately defined general guidelines for what is appropriate or "fair game" in manuscript reviews. The absence of an agreed-upon process thus frustrates us and weakens scientific progress, which depends on the timely release of reasonably defined experimental results for others to explore further. Here, we propose possible solutions to these issues through collective adoption of a community-defined evaluation format and associated standards: (1) a universal template that provides a means by which reviewers and authors can distinguish critiques related to "quality" from those related to "perceived impact" in a way that should provide portability across journals and/or tiers of journals and (2) scientist-driven standards of what a review should not do, providing authors and editors with sensible reasons to dismiss capricious or poorly thought-out criticism.

Benefits of Peer Review

Peer review has many benefits. For example, it provides additional "eyes" that can help identify technical issues that lead to mistaken conclusions, and it ensures that field-specific standards are applied (e.g., the performance and reporting of replicate experiments and application of statistical tests). These types of functions limit the extent to which our field wastes energy or resources by limiting

undue focus upon fully unsubstantiated statements and results that for one reason or another cannot be replicated. Reviewers can also suggest alternative interpretations of the data, leading to alterations in discussion. Ideally, this latter function would rarely lead to an outright delay or blockade of publication, although both a journal and an author will best be wary of publishing work where alternative interpretations are far stronger than the authors' favored conclusions.

Peer review also has taken on a role in prioritizing work to be published in the most prestigious journals. Most favorably viewed, this helps the field and others to direct attention toward ideas and data that are likely to have long-term importance or dramatically advance the field. Assessments of the possible importance of a result can thus place a paper in a journal; that then helps readers evaluate and choose which of the thousands of papers published each week deserve their first attention. In this way, peer review does serve as the first step of evaluation in a contest of ideas. This function is certainly the most troublesome and difficult; it is highly subjective and toys with scientific egos, since it forces the reviewer to credit the author with something that not all studies of equivalent effort will produce. It represents the opinion of a small collection of scientists. Journal editors use our opinions in this area to attempt to provide high-quality content that brings readers back to their journal each week. (Somewhat ironically, we send elite journals our most cutting-edge work in spite of their often longer review processes that can delay publication.) It is arguably the delays associated with the application of our most subjective metrics—the performance of reviewer experiments—that most magnifies the delays for the bulk of the community to share in the work we consider most exciting.

Principles of Peer Review

A collection of 65 mid-career immunologists recently gathered to discuss the future of our discipline. Among these, not a single one had been given unambiguous or consistent descriptions by mentors or through other guidance as to what a peer review should entail. Although journal editors have proposed helpful comments to us, such as to perform reviews in a timely manner (NCB, 2017), they are not necessarily complete in scope nor coming from the right source. As scientists, it is us whose lives and discipline are most impacted by the peer-review process—as readers and as authors—and we provide the service of peer review. It is therefore our actions that most need correction and a higher degree of principled discipline guided by us.

Here, we propose a first iteration of a universal principled review (UP) template, which would provide principles of feedback such as field-specific standards for data quality, completeness and reproducibility, and fair discussion of meaning. Like experiments we undertake in the lab, this one can be assessed and modified in future iterations. Its value may well be in the tens of millions of our own hours spent on peer review, as well as in the success and speed with which our findings are converted into real-world benefits. A primary principle of a template for standards seeks the idea of reciprocity in the golden rule—that we define the kinds of rules for publication that we would like others to apply to our work. A universal scoresheet for review will ideally capture all of the principles above and provide a transparent message of a reviewer's stance on the "publish-ability" of a manuscript and its scientific value. It will also make simpler the assessment of cost-benefit of any additional experiments.

A Universal Framework for Reviews Based on Providing Interpretable, Fair, and Addressable Feedback

We propose six categories as the basis for consideration and outline, along with some of their details, in Box 1.

For the first three "Quality" considerations ("Experiments," "Completeness," and "Reproducibility"), we propose a three-point scale: "1" or possibly "2" would be considered necessary to publish in any journal, with "3" being unacceptable. Quality considerations represent questions of data integrity and logic whose scores should be similar regardless of the journal where a paper is being considered. These are meant to encompass the fundamental scientific demand of rigor.

A fourth "Quality" category ("Scholarship") contains what are sometimes called minor critiques and should never be a substantial basis for a decision to publish, as they largely should be addressed with strong editing. We all recognize that scholarship, however, clearly affects how a work is perceived and how quickly it is accepted.

The last two evaluations ("Novelty" and "Extensibility," together sometimes called "Impact") are more subjective. Acceptable scores for publication may be journal specific; we imagine a score of "1" being consistent with a general-interest (highest-tier) journal, "2" being a top-tier journal within a discipline, and "3" being a highly field-specific journal and an additional category of "4" being added, representing very limited potential readership. A journal-agnostic score for these are clearly the reviewer's opinion and would be expected to vary the most between reviewers.

Box 1. UP Review Categories

OBJECTIVE CRITERIA (QUALITY)

1 Quality: Experiments (1–3 scale)
- Figure by figure, do experiments, as performed, have the proper controls?
- Are specific analyses performed using methods that are consistent with answering the specific question?
- Is there the appropriate technical expertise in the collection and analysis of data presented?
- Do analyses use the best-possible (most unambiguous) available methods quantified via appropriate statistical comparisons?
- Are controls or experimental foundations consistent with established findings in the field? A review that raises concerns regarding inconsistency with widely reproduced observations should list at least two examples in the literature of such results. Addressing this question may occasionally require a supplemental figure that, for example, re-graphs multi-axis data from the primary figure using established axes or gating strategies to demonstrate how results in this paper line up with established understandings. It should not be necessary to defend exactly why these may be different from established truths, although doing so may increase the impact of the study and discussion of discrepancies is an important aspect of scholarship.

2 Quality: Completeness (1–3 scale)
- Does the collection of experiments and associated analysis of data support the proposed title- and abstract-level conclusions? Typically, the major (title- or abstract-level) conclusions are expected to be supported by at least two experimental systems.
- Are there experiments or analyses that have not been performed but if "true" would disprove the conclusion (sometimes considered a fatal flaw in the study)? In some cases, a reviewer may propose an alternative conclusion and abstract that is clearly defensible with the experiments as presented, and one solution to "completeness" here should always be to temper an abstract or remove a conclusion and to discuss this alternative in the discussion section.

3 Quality: Reproducibility (1–3 scale)
- Figure by figure, were experiments repeated per a standard of 3× repeats or 5 mice per cohort, etc.?
- Is there sufficient raw data presented to assess rigor of the analysis?
- Are methods for experimentation and analysis adequately outlined to permit reproducibility?
- If a "discovery" dataset is used, has a "validation" cohort been assessed and/or has the issue of false discovery been addressed?

4 Quality: Scholarship (1–4 scale but generally not the basis for acceptance or rejection)
- Has the author cited and discussed the merits of the relevant data that would argue against their conclusion?
- Has the author cited and/or discussed the important works that are consistent with their conclusion and that a reader should be especially familiar when considering the work?
- Specific (helpful) comments on grammar, diction, paper structure, or data presentation (e.g., change a graph style or color scheme) go in this section, but scores in this area should not to be significant bases for decisions.

MORE SUBJECTIVE CRITERIA (IMPACT)

5 Impact: Novelty/Fundamental and Broad Interest (1–4 scale)
- A score here should be accompanied by a statement delineating the most interesting and/or important conceptual finding(s), as they stand right now with the current scope of the paper. A "1" would be expected to be understood for the importance by a layperson but would also be of top interest (have lasting impact) on the field.
- How big of an advance would you consider the findings to be if fully supported but not extended? It would be appropriate to cite literature to provide context for evaluating the advance. However, great care must be taken to avoid exaggerating what is known comparing these findings to the current dogma (see Box 2). Citations (figure by figure) are essential here.

6 Impact: Extensibility (1–4 or N/A scale)
- Has an initial result (e.g., of a paradigm in a cell line) been extended to be shown (or implicated) to be important in a bigger scheme (e.g., in animals or in a human cohort)?
- This criterion is only valuable as a scoring parameter if it is present, indicated by the N/A option if it simply doesn't apply. The extent to which this is necessary for a result to be considered of value is important. It should be explicitly discussed by a reviewer why it would be required. What work (scope and expected time) and/or discussion would improve this score, and what would this improvement add to the conclusions of the study? Care should be taken to avoid casually suggesting experiments of great cost (e.g., "repeat a mouse-based experiment in humans") and difficulty that merely confirm but do not extend (see Bad Behaviors, Box 2).

"Experiments," "Completeness," "Reproducibility," and "Scholarship" scores are absolute values based on best practices and are meant to be the least subjective. "Extensibility" and "Novelty" represent the more subjective measures and are judgements in the contest of ideas, and this is overtly acknowledged. A journal editor will of course have the final say in deciding how that score will translate into acceptance or rejection. However, scoring using this rubric would always be universal—whether one's peers consider the quality acceptable should not vary according to which journal solicited the review.

A transparent adoption of these principles suggests a template similar to those from NIH study sections where strengths and weaknesses are called out. Importantly, where weaknesses are found, paragraphs that follow from the score must be specific. Additionally, if we as reviewers propose experiments or changes in text, we should be required to provide our own assessment as to how an experiment will increase the score in this category and how long it will take. Generally, if the score is short of a "1," reviews must indicate what is necessary to get the score higher and estimate its timing,

Box 2. Identifiable Bad Behaviors

- Blanket statements that lack justification (e.g., "I don't believe the data") indicate that the reviewer is not an active reader or a careful judge. A review should call out concrete experiments, controls, etc. to address their issues. If a modification is needed for a method, a criticism should call out experiments to the degree that they can describe how they are achievable in a timely manner. A review should seek to pull meaning from the labors of others wherever possible.
- Claims of non-novelty based on papers that are non-synonymous, reviews, or cartoons (not primary data) are bases for eliminating a review or a score in that category. In particular, if a reviewer suggests that something is known, but no clear experiments that definitely show it have been performed, then their comment is false. Due to the proliferation of review articles and journals dedicated to reviews, many ideas may have been conceived but not experimentally demonstrated. The data need to be shown, and such experiments are very valuable. Further, the claims by a reviewer of non-novelty or "known" may not be shared by the broader community. It is fair for a reviewer to suggest that a review or primary article be cited and discussed (under "Scholarship").
- Critiques that compare current data against "ideas" or "works of person X" can suggest inappropriate bias. A paper and its data should be put up against data from the past and not just ideas or weighty individuals. Previous publications should have to be weighed for their data while very carefully scrutinizing the source of momentum of a possibly incorrect or incomplete idea. The relevant figure from previous studies (e.g., "PubMedID XX Figure Y") that seems to propose alternative conclusions should always be explicitly called out. The conclusions we hold dear from past papers might be based on data or experimental approaches that would not meet our current standards. We cannot rely on our recollection of our take-home message of an old paper. We have to go look at it and decide if we would reach the same conclusion or if the interpretation an overreach or may be tempered given what is now known.
- A reviewer who significantly confuses extensibility-level questions with quality and reproducibility and uses this as a primary reason for low scores in the "Quality" categories should be eliminated. A good paper opens many new questions, but these new questions are often more appropriate for the next paper(s). When using the proposed guidelines, asking for further experiments (extending the result, possibly adding import) should be distinguished from assessing a feature that informs "Quality," like the adequate repeat of an experiment. A reviewer should beware if they propose enough experiments for an entirely new (and entirely different) paper. What does this paper show, and does it support the fundamental idea of the publication regardless of whether different experiments might also address that? Likewise, a reviewer should not delay publication of mouse studies to demand data regarding human trials. Doing such significantly undervalues the value of basic science.
- A reviewer who shows evidence that their ultimate satisfaction is the primary goal. This is most evidenced by a reviewer following a response to their primary review by asking for new material or material that was not a direct consequence of a point raised in an initial review. Rather than increasing the value of science, this behavior mostly adds further burden in the whole process and delays science. Unless material evidence of fraud has emerged (or something equally onerous), reviewers should raise all major quality and impact issues in the first review. A general guideline should be for a reviewer to simply respond yes or no as to whether an author's increased discussion, modification of text, or added experiment addresses the questions raised in a primary review.

typically within 1 month, 3 months, or beyond. Anything beyond 1 year should not be proposed or indicated as such. The goal of this is to indicate how this work (not another paper entirely) would be suitable for publication in any journal, if that is possible. In the interest of transparency, we believe that all comments should be to the author directly; no scientific comments should be hidden in confidential comments to editors.

A Series of Guidelines to Identify Science-Disabling Reviews

A corollary to the need for principles is that malfeasance by reviewers needs to be clearly defined by our community so that it may be minimized or neutralized. To be fair, when queried, all authors of this article acknowledged to having been sometimes lazy or snarky when serving as reviewers. Without published norms, this will happen more often. We suggest the need to reverse a trend whereby some reviewers appear to try to act as the gatekeepers or owners of ideas. Our greatest value is to ensure overall quality. Too many papers are rejected based on a conflict of concepts. We posit that it should be the duty of broad readership, performing experiments over time rather than precedence, that quickly determines winners in the contest of ideas. It is fair to point out where an experiment fails to support a conclusion ("Quality") or where an idea fails to be a sufficient advance, but the latter in particular needs solid documentation to support the viewpoint.

This is an area for the golden rule, and there seem to be clearly identifiable reasons to fire a reviewer and dismiss poor criticism. We outline a few of these on which we all agree in Box 2.

As a collective, we argue that an author, on seeing reviews with those features, should have the right to request that a reviewer be replaced ("We submit that Critique #3 by Reviewer #1 is in violation of our community's UP guidelines. We respectfully decline to address some/all aspects of this critique in our revised manuscript."). Alternatively, editors would reconnect with the reviewer to seek an improved critique. When journals share completed reviews among the collection of reviewers, as they should, bad behaviors should be called out (and of course, this also provides the ability of reviewers to concur or "rescue" comments that are actually meritorious). We would hope that editors, on observing a scientist writing this kind of text, should quickly cease to solicit reviews from these individuals.

In short, there needs to be benchmarks that everyone agrees upon, given the importance of letting ideas and associated data loose for broad consideration.

Implementation: Integrating with the Publishing Machinery

Given the large number of journals in a field, experimenting with UP does not initially require buy-in from all journals, although we would hope that editors will take the community's needs into account. Box 1 represents a template that can be used in any journal—we will simply paste the text into the author's text box. As authors of this document, each of us is

agreeing (1) to try UP categories in their reviews over the next year and (2) to support editors who act upon the bad-behavior exclusions of UP.

For their part, journals could use our design to set realistic and clear expectations of what they emphasize and demand. For example, a top-tier immunology-specific journal might suggest: "Our journal requires average 'Quality' scores of 1.5, but 'Extensibility' can be $\leq$3." They can also state their willingness to remove reviewers and/or request changes when significant poor behavior is witnessed, as described above.

We also note that the nature of scientific publishing itself is in flux with the useful appearance of pre-publication websites that allow first glimpses of data pre-peer review. UP, having an element of portability, could allow a model wherein scientific societies move the reviewing process entirely away from journal-specific administration and can allow journals to "bid," including for papers that are already being more widely considered on a pre-publication website. Journals could make the effort to transfer the review to the subsequent journal and share the identity of the reviewer on demand and with their assent. Different journals might also list discrete rules as to whether three total positive reviews are required as compared to concurrence of the first three, which is a current norm. If a journal requests the first three for concurrence, we believe they should allow a large number of reviewers to be excluded by the author. Finally, an UP review needn't only apply to papers at the time of peer review and could also be applied to historical manuscripts, should our community devise a mechanism for posting these in an accessible and meaningful way.

Concluding Remarks

All who have ever acted as mentors know well the phenotype of the student who, afraid to fail in his or her first experiment in a series, simply does no experiments and continues to only read journal papers and contemplate the possibilities. We all know full well that no advances are made without jumping in. In proposing that our community adopt and ultimately refine a framework like UP, we are well aware that iterations will be required. First experiments may work in some areas but not in others, and unintended consequences may well result. This platform of principles will not guarantee that the work you consider your best makes it through to the highest-impact journals. However, we submit it as an improvement on the status quo, and we as a community will have opportunities to refine this in subsequent iterations. Obtaining metrics to guide the process, just as in any experiment, is key. A few strong journals could help collect data to serve their readership. The primary guide for this should not be impact factor, an imprecise measure, but simply the time to publication experienced by investigators and the overall ease of the process. We cannot know the result of changes such as these outlined here unless we try, and as a small community, we hope that others will join us in improving our discipline's dissemination of important results and ideas and embrace this experiment. In the meantime, we believe the following should be our ethos on review:

- There is no need to "guard the field" but rather to provide balance and ensure the dissemination of high-quality experiments that provide sufficient detail for reproduction.
- We all need to train the next generation to practice peer review in a responsible way during journal clubs. Too many training institutions teach that a critique should emphasize weaknesses; include trainees in your process by asking them to help when appropriate and mentoring them on principles from the beginning of the process to the end.
- Review criteria will not be one-size-fits-all. What we propose above seems to us a reasonable framework by which to operate. We welcome future revisions or better alternatives.
- Finally, and most simply, write each review as if you will sign your name next to it.

We hope this attempt to mildly codify our peer-review process will improve the entire scientific endeavor. For those interested in signaling public assent with some or all of these ideas, please visit the UP review page at https://immunox.ucsf.edu/future-immunology.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.11.029.

ACKNOWLEDGMENTS

These ideas came out of discourse at a meeting of immunologists in Skamania, WA in June 2019 (Immuno-Skamania) organized by Mark Ansel, Ananda Goldrath, Max Krummel, and Marion Pepper. We thank all attendees, Burroughs Wellcome Trust, and UCSF ImmunoX Initiative for funding to help defray the costs of the meeting. We thank Vincent Chan and Isabelle Tingin for assistance in preparing the manuscript.

REFERENCES

Kaelin, W.G., Jr. (2017). Publish houses of brick, not mansions of straw. Nature *545*, 387.

NCB (2017). Principles of refereeing. Nat. Cell Biol. *19*, 1005.

Ploegh, H. (2011). End the wasteful tyranny of reviewer experiments. Nature *472*, 391.

Rubriq. (2013). Peer Review: How We Found 15 Million Hours of Lost Time. https://www.aje.com/arc/peer-review-process-15-million-hours-lost-time/.