Response to reviews

Manuscript title: Finding the right words to evaluate research: An empirical appraisal of *eLife*'s assessment vocabulary.

We are grateful to the editor and reviewers for their rigorous assessment of the manuscript and thoughtful comments. Below we provide a point-by-point response to the reviews and note any corresponding changes we have made to the manuscript. All changes are also highlighted in the latest version of the manuscript using 'track changes'. We have numbered editor/reviewer comments sequentially. Reviewer comments appear in black and our responses are in blue.

Editor comments

1. This Stage 1 Registered Report proposal details a study which aims to address in interesting metascientific issue: that of linguistic ambiguity in the journal eLife's assessments of manuscripts under its new model of eliminating accept/reject decisions after peer review. The authors of the proposed study point out that the model's success will partly depend on how clearly with prospective readers. They argue that, at present, some of the wording contained in eLife's manuscript assessments is counterintuitive and ambiguous. The authors have designed a study to explore whether the language used in the eLife assessments will be interpreted 'as intended' by readers.

   I received four thorough and constructive reviews of this proposal and have used those to supplement my own thoughts and assessment of this proposal. In my opinion, the proposed study has potential to make a useful contribution to the metascience field, as well as being a valuable source of information for other journals potentially interested in following the novel path made by eLife. That said I have some concerns about validity and generalizability, as well as about the assumptions of the study.

Author Response: Thank you for the helpful comments.

2. First, I think it might benefit the manuscript if the authors motivated the study a little more. To be clear, I think it's important that the wording of these assessments is questioned, but I also think that issues with interpretation will be present any time we use a qualitative descriptor. Is it at all possible to use words that will not carry with them some variance in interpretation, especially across a population with varying proficiency and understanding of English? I'm not convinced that other wording would necessarily bring fewer varying interpretations with it. As one reviewer commented, do we in fact know that others (i.e., outside of the author group) find the wording ambiguous to the degree that it would undermine the assessment in question? I suggest that the authors argue a little more as to why the wording is problematic (including indications that it indeed is, if there are any), and motivate why their choice of alternatives would be better. I realize such arguments are already present in the proposal, but I have to admit I don't find them as compelling as I think they could be.

Author Response: We agree that all words have the potential to be interpreted differently by different people; however, words are not equally ambiguous or overlapping in meaning. It is therefore possible to improve accuracy of communication by choosing better words (for empirical examples, see the studies we cite regarding the interpretation of probabilistic language). As to whether others beyond our author team find the eLife vocabulary problematic and the alternative vocabulary to be an improvement, we feel we have made the case as strongly as we can based on non-empirical arguments alone — ultimately these are empirical questions and it's the purpose of the proposed study to address them.

3. Second, and this is something that reviewers also pointed out – have eLife been asked about their intentions regarding the language they used? Indeed, to properly assess whether wording is being perceived as intended, we need to know what that 'intended' actually entails. As one reviewer mentioned, eLife might be intentionally using ambiguous language. Without asking, we do not know this. I recommend that the

authors confer with eLife to clarify this rather than relying on (what appear to be?) assumptions about what eLife intended or did not intend to convey with their wording.

Author Response: Yes, we mentioned our contact with eLife in the cover letter. We wrote to Damian Pattinson (Executive Director of eLife) and Michael Eisen (Editor-in-Chief of eLife) on 5th May 2023 and shared the study protocol with them. Tom Hardwicke then met with Damian in person on 8th May 2023 at a conference dinner in Washington D.C. and discussed the proposed study. We don't mention these conversations in the manuscript because they were informal and therefore we presume off-the-record. The clearest public articulation of eLife's rationale appears in their article introducing the vocabulary: "…to help convey the views of the editor and the reviewers in a clear and consistent manner, we have created a common vocabulary…" (available at https://shorturl.at/hAPT4). We've added this quote to the manuscript. If our empirical results find considerable variation in the perception of the eLife vocabulary, this would be incompatible with the goal of "clear and consistent" communication.

4. Third, I am, along with one reviewer, concerned about the sample. The authors plan to use this study to assess how language is perceived, which implies that the conclusions will strongly depend on those perceptions. The perceptions will strongly depend on the sample. In the proposal, the description of the sample boils down to convenience (this is made clear by the authors). This is problematic, in my opinion. If the findings of the proposed study are to hold any real validity beyond a small segment of largely young, white, middle-class English-speakers (i.e., university undergraduates, likely mostly from Australia), the sample must be actively constructed to include some demographic variety. This is especially relevant when one considers the readership of eLife. That is, they will represent a much larger proportion of the population than the sample the authors plan on drawing their data from. The authors suggest that this is likely non-expert readers, but their 'non-expert' status is only one element of the sample to be considered. One reviewer suggests stratifying to include a wide variety of people, and I think that's a good idea. In my

> opinion, if the authors decline to take this into account in their design (for feasibility reasons, which are certainly understandable) I don't think the resulting findings will be nearly as helpful as they could be. The authors do note this in the limitations section, but I think they understate it, if I'm being frank.

Author Response: We agree that a representative sample would be ideal; however, it is unclear how to achieve this in practice. The main issue is that we do not know the demographic characteristics of the relevant population (potential readers of eLife or scientific journals more broadly), so we cannot attempt to emulate those characteristics in our sample. We note that it is not necessarily the case that interpretations will 'strongly depend on the sample' — that is unknown. We should also clarify that our sample will not consist of university undergraduates from Australia — we are recruiting from the Prolific platform which recruits users from the general population across a range of countries (though there is heavy over-representation in the UK/US). We have now added the following to the manuscript so readers have a bit more information about Prolific's participant pool:

"As of 23rd August, 2023, the platform has 123,064 members. Complete demographic information about the members is not available as demographic screening questions are voluntary. Based on the available responses, 30% of Prolific members say they are aged between 18-25, 58% say they are aged between 26-50, and 12% say they aged between 51-100; asked about their gender, 35% identified as a man, 46% identified as a woman, and 2% identified as non-binary; 18% say they are currently a student; 87% say they are fluent English speakers; 31% say they have UK nationality and 31% say they have USA nationality; when asked to report the highest level of education they have completed, 25% said undergraduate degree, 12% said graduate degree, and 2% responded doctorate degree."

We anticipate that the most relevant demographic characteristics are education status and language. It seems reasonable to assume that most eLife readers are highly educated (because the content is technical) and can speak English (because the content is in English). On these grounds, it seems reasonable to start this line of investigation with a convenience sample that aligns with these characteristics. If our sample of highly educated, fluent English speakers have difficulty interpreting the eLife vocabulary, then we expect others are likely to have even more difficulty. We do not deny that other demographic characteristics are potentially relevant or interesting, but investigating them is beyond our scope/resources.

After reflecting on this issue, we have decided to modify our inclusion criteria so we are recruiting individuals with more advanced educational qualifications. We anticipate that this will better represent people who are likely to read eLife articles. The inclusion criteria section now reads as follows:

“Participants must have a >= 95% approval rate for prior participation on the recruitment platform (Prolific). Additionally, Prolific pre-screening questions can be used to ensure that the study is only available to participants who meet certain criteria. We will only recruit participants who report that they speak fluent English and are aged between 18-70 years. Additionally, participants must report that they have completed a doctorate degree (PhD/other). If we do not reach our target sample size within 3 weeks, we will expand recruitment to individuals who report that they have completed a graduate degree (MA/MSc/MPhil/other). If we do not achieve our target sample size after an additional 3 weeks, we will expand recruitment to individuals who report that they have completed an undergraduate degree (BA/BSc/other).”

Additionally, we have modified the relevant limitation statement in the discussion section:

“Though we wish to understand how *eLife* readers interpret the vocabularies of interest, it is unclear how to recruit a representative sample without information about the demographic characteristics of *eLife* readers. We anticipate that the most relevant demographic characteristics are education status (because the content is technical) and language (because the content is in English). We are therefore aiming to recruit people with advanced educational qualifications (preferably doctoral degrees) who speak fluent English. Relative to our sample, we expect that *eLife* readers are probably more likely to be professional scientists working specifically in the life sciences, with some, but not necessarily fluent competency with English. These differences may impact the generalizability of the results. Note however, that *eLife* explicitly states that *eLife* assessments are intended to be accessible to non-expert readers."

5. Finally, one reviewer mentions a confound that I think the authors might attempt to address in some fashion. He points out that while most readers of the eLife assessments will only see one or two assessment reports, the study's participants will see multiple. He comments that the experiment might not sufficiently simulate what goes on in real life. I agree with this. Do the authors have any ideas as to how to get around this, or a substantive argument as for why this shouldn't undermine the usefulness of the eventual findings?

Author Response: We are grateful to the reviewer for highlighting this issue. We agree that this part of the experiment may not reflect readers' natural experience with *eLife* assessments — we expect some readers may be interested in drawing comparisons between vocabulary used in different *eLife* assessments, but some readers will probably only read one or a few eLife assessments, at least within a short period of time. Thus, the specific methodological concern is that by being exposed to vocabulary phrases in close temporal proximity, participants may be prompted to make comparisons between them, which they will not necessarily be doing in natural interactions with *eLife* assessments. Additionally, participants' judgment of each phrase may be impacted by their judgment of prior phrases (i.e., carry-over effects).

An alternative design choice that addresses these issues would be to show each participant a single phrase only; however, this would require an untenable sample size. Instead, we propose to disrupt participant's efforts to make comparative judgments as best we can. Currently, comparative judgements and carry-over effects are at least partly addressed by two measures (1) participants cannot go back and edit prior responses, so any inter-trial influence can only be uni-directional; and (2) the order in which phrases are presented is random, which should minimize any systematic influence. We also have decided to add a third measure (3) a 'wash-out' period between each trial. Participants will now complete a short (15 second) filler task (simple multiplication problems) in between each trial in an effort to stop them ruminating on their judgment on the prior trial. We have added the following to the procedure section of the manuscript:

“After each statement, there will be a 15 second filler task during which participants are asked to complete as many multiplication problems (e.g., 5 x 7 = ?) as they can from a list of 10. The multiplication problems will be randomly generated every time they appear using the Qualtrics software. Only numbers between 1 and 15 will be used so most of the problems will be relatively straightforward to solve.”

6. Other comments by the reviewers should also be taken into consideration by the authors, either resulting in changes to the protocol or arguments for why they may be ignored. I hope the authors find these points reasonable and can implement them or assuage the concerns I and the reviewers have. I look forward to reading a revised protocol!

Author Response: We appreciate the helpful comments, thank you!

Reviewer #1 Comments

7. Reviewed by Chris Hartgerink, 06 Jul 2023 12:03

Thank you for the invitation to review the Stage 1 report for "Finding the right words to evaluate research: An empirical appraisal of eLife’s assessment vocabulary." It was a pleasure to read this report. It's a fantastic proposal, and I only have minor points to make.

What went well

This stage 1 report was encompassing and provided a fantastic introduction to the material. I've followed eLife's changes, but was not aware of the details that form the basis of this paper. I felt like the authors shared their expertise generously and that I learned why this work is important. It's of course fantastic to see a paper on communication be this strong in communicating itself.

I also very much enjoyed the dynamic document available to highlight what analyses will be included. I wish every manuscript that crosses my inbox was as rigorous as this. It made reviewing a breeze and I felt confident in understanding what's going to happen.

Author Response: Thank you! And so happy to hear you enjoyed the dynamic document!

8. What could be better

My only minor connotations are the following:

This is called a "psychometric study" - that's a bit of a stretch given that there is no psychometric modelling happening. I'd suggest to drop the psychometric altogether.

Author Response: Thank you, we agree and have removed the description ‘psychometric study’ and replaced it with ‘experiment’. We have also updated the text: “In this study, we intend to empirically evaluate the psychometric properties of the eLife vocabulary (Table 1) and assess whether an alternative vocabulary (Table 2) has more desirable properties" to “In this study, we intend to empirically evaluate how the *eLife* vocabulary (Table 1) is interpreted and assess whether an alternative vocabulary (Table 2) elicits more desirable interpretations”. We have also updated “Our study is modelled on psychometric studies” to “Our study is modelled on prior studies”.

9. The introduction highlights the potential confusion around the current vocabulary eLife uses - is there any anecdotal evidence that indicates confusion goes beyond the author team? I read the references, and I think that's sufficient evidence to take this step, but it would be good to know whether in this specific instance that confusion is already observed

Author Response: We’re not aware of any relevant anecdotal evidence.

10. Continuing on that note - it might be helpful to include some context on why this specific eLife vocabulary was included to begin with. I imagine they deliberately chose this, and had certain considerations. It would strengthen the final report to have this, and if that is not already public, it might be worth reaching out to Mike Eisen (or someone else from eLife) for some comments on the vocabulary. This will help ground the reader and may provide you with worthwhile information and stepping stones to potentially improve the language in practice at eLife (if that's something you'd like)

Author Response: The only rationale for these phrases we have seen is that they are based on a “set of widely-used expressions from the summaries written to date” (in https://shorturl.at/hAPT4). We have been in touch with eLife (see our response to comment 3), but have no additional information from them on this particular point.

11. What to watch out for

There is no indication that ethics approval has been granted or is not a requirement for this study. I do not see any obstacles, but nowadays there are (varying) obligations that all human participant studies require ethics approval. It would be a shame for a study like this to get stuck on a procedural note later in the process.

Author Response: The ethics statement on page one states that the study has ethics approval.

Reviewer #2 Comments

12. Reviewed by Veli-Matti Karhulahti, 26 Jun 2023 11:28

I’m excited to review this MS, as it tackles a meta-scientific issue toward which I have great personal curiosity and interest. The work is clear and well-organized. I’m not an expert in statistics so I leave in-depth commentary on that area to other reviewers and recommenders (with a few exceptions). I’m an interdisciplinary researcher with a background in philosophy/theory and qualitative work; this info hopefully helps interpreting my feedback. The comments come in no particular order but I list them to make it all easier to read.

Author Response: Thank you for your helpful comments!

13. The premise of the MS is as follows: “Our understanding (based on eLife’s New Model, 2022) is that eLife intends the common vocabulary to represent different degrees of each evaluative dimension on an ordinal scale” (p.2). This is later paraphrased on page 5: “The success of eLife assessments will depend (in part) on whether readers interpret the common vocabulary in the manner that eLife intends.” As I read it, the MS is based on the authors’ guess about what “eLife intends” in their recent release. I don’t know the authors’ positionalities (e.g., if some are or have been affiliated to eLife) but to me it would be obvious to *talk* to eLife as a Pilot #1 interview and simply ask what they intended (instead of guessing). This would make the foundations of the study stronger. Open dialogue could correct misunderstandings and save time.

Author Response: Thank you for highlighting this — we have been in touch with eLife (see response to comment 3).

14. Related to the above, the MS correctly points out how the significance and support domains involve many dimensions. The authors mention breadth/scope and degree, but personally I see many more. E.g., a study could be “valuable” for pragmatic reasons in a subfield and immediately save human lives (vaccines etc.) and this feels incomparable to theoretical development or other contributions to cumulative models or science at large. Likewise, “support” seems to conflate various dimensions related to

methodology; e.g., would larger effect sizes or statistical power contribute to “more support”, and how would e.g., qualitative evidence or clinical case studies be assessed in this domain? On the other hand, if “support” here simply refers to a general commitment to established methodology (as is implied by “exemplary use of existing approaches”), this doesn’t seem to have much to do with “support” but rather methodological rigor (e.g., a correctly reported patient case study of N=1 could have “exceptional strength of support”). This goes back to my #1: it would be important to know what the intended areas of evaluation really are -- and whether they are intended to operate purely as continuous ordinal scales -- before developing competing vocabulary.

Author Response: Yes great points; as noted in the manuscript, we hope the simpler alternative vocabulary will reduce this conflation.

15. I really like the Box 1 example, as it demonstrates the practical problems related to the current eLife model. Indeed, while the chosen words now have clear definitions and meaning, the full description additionally uses various undefined evaluations such as “huge amount of data”, “very interesting observations”, “well supported by the data”, and a promise to “sharpen our understanding.” That is, while the reader could pick up the meanings of one or two now-defined terms, most of the semantics remain vague and undefined. If the vocabulary is meant to operate as a continuous ordinal scale, the hermeneutic benefits over simple numeric ratings seem to disappear in actual use. This comment is perhaps directed more to eLife than the authors, but could be useful to reflect on in the MS.

Author Response: We agree with the reviewer that the expanded eLife definitions introduce new words and phrases that are themselves subject to a wide range of interpretations. As we will only test participants’ ratings of the single word labels (not the expanded definitions provided in Table 1), this is not especially relevant to our study, but is a further critique of the eLife approach.

16. It’s not stated who the participants are. It’s noted that the data comes via Qualtrics and “our sample are more likely to be undergraduate students” (p. 13). It would be good to clarify this aspect of the data.

Author Response: Thank you for highlighting this. Qualtrics is used to deliver the study; however, the sample is recruited from a platform called Prolific (see ‘sample source’). We have now added more details about the Prolific platform to the manuscript (see response to comment 4).

17. There’s one hypothesis in the MS, expecting a higher accuracy for the alternative wordings. That’s quite simple, but I have two notes (naturally ignore some of this if it feels not right). First, I think it always helps the reader when hypotheses are framed in such way that: *why* something expected, *what* is expected, and *where* do the results lead us (“Because X, we expect Y. If Y, that’s interesting because Z). Currently, the hypothesis is kind of “hidden” in a wall of text. Second, I would really like to see something in the *where* part (“theory that could be wrong”). It’s written in the final table that this is an applied RQ, but they also have implications and it would be important to spell out these implications at Stage 1 (especially in the final table, which serves as a reference at Stage 2).

Author Response: We have added: “We expect ranking accuracy to be higher for the alternative vocabulary relative to eLife vocabulary because it is designed to be more structured and less ambiguous.“

18. Page 12 says there are no planned analyses for RQ3. Although I strongly support exploratory and descriptive work, I would also like to see a bit more detail here (skip if it doesn’t feel right). Another note that may not be directly relevant: it feels the auxiliary hypotheses behind “To what extent do different phrases used to describe scientific research elicit overlapping interpretations” is that everyday terms are expected to “elicit” certain interpretations, and this is what eLife is trying to catch. However, as I see it, the purpose of the new eLife vocabulary was to give *new technical

definitions* to these terms and not to try catching their organically eliciting universal features (I can be wrong). It could be a useful enterprise for any journal to build a glossary for assessment terms, but that seems to be a different effort vs trying to capture the "true" meaning of everyday terms (see my #8 too).

Author Response: Regarding analyses for RQ3, please see our response to comment 29. Regarding technical definitions — although eLife provides definitions in their article introducing the vocabulary, these definitions are not included in the eLife assessments (see caption of Table 1) — so it seems unlikely that most readers will ever see them.

19. I list a number of small technical comments.

There's only one attention check. I know there are many opinions about this and no single correct solution, but subjectively I like giving participants the opportunity to correct their mistake once with a follow-up attention check if they fail (some people have naturally low attention by neurodiversity). Alternative checks could be used too.

Author Response: Thank you for this suggestion. As noted by the reviewer, there's no single correct or consensus solution to this issue. It's plausible that attention checks can make participants suspicious and change their behaviour, so we don't want to include too many of them. On the other hand, we want to avoid data contamination from participants who are not paying attention. We think a single attention check at the end of data collection is the right balance and at that stage it cannot influence participant's behaviour on the task. Note that performance on the attention check will have no consequences for participants themselves (they still get paid etc).

On reflection, we have decided to add two additional measures to facilitate higher quality data. Firstly, we have added the following inclusion criteria:

“Participants must have a >= 95% approval rate for prior participation on the recruitment platform (Prolific).“

Secondly, we have added a lower-bound to our time exclusion criteria:
“all of a participant’s data will be excluded if (1) they take less than 5 minutes or more than 30 minutes to complete the task”

20. Page 8: “ignoring exclusions” -- for this non-native English speaker it’s not very clear, perhaps just with or without exclusions?

Author Response: Thanks for flagging this — we’ve changed to ‘without exclusions’.

21. In the inclusion criteria, me coming from Finland, I don’t know what “A-levels” are.

Author Response: We’ve changed the educational levels so this is no longer applicable (see response to comment 4).

22. Last, I briefly return to the RQs. As per Wittgenstein, words gain meaning through language games, and one and the same word can have multiple meanings depending on the context. Here too, different use contexts could yield different messages, as “evidence” remains relative to a “claim.” Sadly I’m not a linguist nor semiotician, but I would think there’s existing evidence/theory in these fields to which the present findings could contribute (or learn from). The design nicely borrows from earlier studies on probabilistic statements, but I’m also reminded of e.g., Manfred Krifka’s and Teun van Dijk’s work, which could be informative at least at Stage 2. I don’t have the topic expertise to be able to pinpoint what specific model/theory would be helpful in this context, but I would love to see further interdisciplinary bridges in future investigation. Again, just skip this if it’s not helpful (hopefully another reviewer is a topic expert).

Author Response: Thank you for these suggestions. Our motivation is to address an applied problem rather than test a theory; but we agree that future work on this topic may benefit from collaboration with linguists.

23. Although significance and support are rated independently, would it really be possible for a paper to have e.g., “landmark” significance and “inadequate” support? (Feeling the future?) In reality it seems the two domains are somewhat artificial and not necessarily worth separating, and if so, it could be worth rethinking how to frame this study to maximize its application value for the academic world and its evolving journals.

Author Response: We agree. The reviewer may be interested in the eLife assessment for this recent paper (https://elifesciences.org/reviewed-preprints/89106#assessment) which says that the study “would be a **landmark** finding. However, the evidence for these claims is considered **inadequate**”.

24. I hope some of my comments are useful; just ignore those that aren’t. If something feels unclear or unfair, I can be contacted directly. Best wishes for revising this important paper that will likely yield compelling evidence,

Veli-Matti Karhulahti

Author Response: Thank you!

Reviewer #3 Comments

25. Reviewed by Štěpán Bahník, 07 Jul 2023 19:23

The study aims to assess the perception of vocabulary describing studies used by eLife and an alternative set of descriptions which is supposed to improve on the descriptions used by eLife. The study is practical and well designed. I have just a few comments:

Author Response: Thank you for the helpful comments.

26. Readers who see a description of a study usually probably read just one description used for a single study. The experiment, however, has the participants evaluate the whole set of descriptions as well as its alternatives. The comparison of the different descriptions is thus much easier and more likely. It is possible that one set of descriptions is easier to order while the other is easier to interpret when seen alone. As an absurd example, it is possible to imagine vocabulary “100% support”, “75% support”, etc., which would be ordered correctly, everyone would rate the support similarly on the scale used in the study, but it would be probably useless in practice because it would be hard to know what the percentages mean when taken out of the context of the current study.

Author Response: Please see our response to comment 5.

27. The strength of support has 6 levels in the eLife vocabulary, but just 5 levels in the proposed one. How is this taken into account in the analysis?

Author Response: Thanks for prompting us to think about this issue. We don’t think much is needed to address it because our analyses are largely descriptive and address an applied problem. So although the larger number of phrases for the eLife support dimension perhaps puts it at a small disadvantage (e.g., because there are more opportunities to misrank) relative to the alternative vocabulary, that is just the way things are.

One change that is needed however, is to ensure that for the Kendall distance (Kd) analysis, we use the normalized distance instead of the raw distance so that Kds for the different vocabularies are comparable. We had not previously mentioned this in the protocol, so we have added “We will report the normalized *Kd* because one vocabulary set has six phrases and the other sets have five.”

28. The answer scales should probably show percentages, otherwise people might not know how to interpret them (it is mentioned just once in the instructions that the participants should answer in percentages). Another question related to the scales is whether they even make sense. That is,

what does it mean that a study is 80% important or that it provides 30% strength of support? I would have little idea what these statements mean. It is possible that this will not be a problem if people use the scales consistently, but that seems like something that should be established or at least discussed.

Author Response: Thanks for spotting that — we've added percentage symbols. As for the deeper issue of what percentage ratings mean — terrific question! Our response is more pragmatic than philosophical. We agree that statements like "30% strength of support" sound a little odd and might not survive a deep conceptual analysis; indeed, that is partly why we advocate for an alternative vocabulary consisting of words rather than numbers. Our use of percentages is to measure participant's interpretations, rather than measure or communicate importance / strength of support. We expect that participants will be able to use a natural interpretation of percentages in terms of "an amount of something".

29. It should be possible to come up with a way to analyze the research aim 3. For example, the ideal scale should probably be interpreted to have the subsequent levels equidistant and cover the whole scale (0-25-50-75-100). It is then possible to use a difference between this ideal interpretation and the participants' interpretations to assess the vocabularies. Similarly, for research aim 1, it is mentioned what will be reported, but not what will be then interpreted. In the simulation, IQR is used for interpretation, but it is not mentioned in the article itself.

Author Response: We agree with the reviewer that there are statistical tests that could be run on the data to quantify overlap and spread in responses to different items. However, we do not think there is a clear, quantitatively specifiable "ideal" spread or distribution. For example, if the distribution of responses to the different items were largely non-overlapping, but only covered the response scale from 30 to 100, it is not clear that this would be suboptimal. It could simply reflect that respondents' subjective interpretation of a 50 out of 100 on the response scale is something like "terrible" (similar to the US marking system, or to most online rating systems such as reputation scores on Uber or Airbnb, where scores are clustered at the upper end of the scale).

For both research aim 1 and research aim 3, we do not have specific quantitative benchmarks and so our interpretation is admittedly under-specified. We think this is preferable to committing to quantitative benchmarks given how little is known about what these distributions might look like.

Regarding the IQR, we do mention it in our description of our analyses for research aim 1.

30. Is it correct that the example on p. 19 yields an odds ratio of 2? It seems to me that the odds would be 2, but odds ratio would be (0.2/0.8)/(0.1/0.9) = 0.25/0.11 = 2.27. But, I might be mistaken.

Author Response: Thanks for checking. The reported odds ratio is a ‘McNemar odds ratio’ which we believe is more appropriate for a repeated measures design. In this case, this is the ratio of participants’ whose eLife rankings did not match and alternative rankings did match (60) and participants whose eLife rankings did match and alternative rankings did not match (30). Thus, the McNemar odds ratio in this case is 60/30 = 2. In the manuscript we have now clarified this by saying ‘McNemar odds ratio’ rather than just ‘odds ratio’.

31. Signed,

Štěpán Bahník

Author Response: Thank you for the helpful comments!

Reviewer #4 Comments

32. Reviewed by Ross Mounce, 04 Jul 2023 13:06

Comments on the introduction

Perhaps I am old-fashioned but it might have been nice to turn-on line-numbering on your manuscript to make it easier for reviewers to refer to specific lines. One for next time perhaps :)

Author Response: We've added line numbers.

33. I think it would be useful to mention to readers that making peer review reports publicly available alongside the published paper, on a MANDATORY basis (for all research papers in that journal), has been practiced at some journals for over TWENTY years now. Examples include BMC Medicine since the start in 2003 & Atmospheric Chemistry and Physics . The intention of this is not to promote any particular brand or publisher, but rather to help the readers of this research understand that in some communities this practice is very far from 'new' or 'untested', it is NOT a new innovation. Publicly available peer review reports are a tried and tested system.

Merely stating a "growing number of journals" is true but in my opinion is insufficient context.

I would also quibble with representing the entirety of the new model eLife has chosen as to "go even further". Specifically, abandoning 'accept/reject' is not on the same plane as publishing the peer review reports. By abandoning 'accept/reject', eLife are not going further in the direction of transparency, but rather they travel further on a different plane, that of egalitarianism(?)

"These improvements" - I'm not an eLife basher. But is it not a little bit subjective to proclaim the changes are definitely improvements? I'd just call them changes to the eLife publishing model/process for the sake of objectivity.

"their success will depend on accurate communication with readers" hmmm… this is a little bit debatable. What do you, the authors of this manuscript define as 'success' for eLife? What does the eLife leadership team define as ‘success’ for eLife? What do authors who have published with eLife both before and after the changes see as ‘success’ for eLife?

Success if mentioned at all needs to be defined. Success may be measured or assessed differently by different stakeholders.

It seems quite a bold assertion to say that success (however defined) depends on accurate communication (of what exactly?) with readers. I would say there have been many journals that have been commercially successful despite having very poor communication with readers. Likewise, taken from a different viewpoint of ‘success’ there are lots of small not for profit run journals that have been tremendously successful from the point-of-view of publishing rigorous reproducible robust research, consistently highly relevant and thought-provoking to those interested in a small discipline – these would all be considered ‘failures’ from a narrow financial profit/surplus perspective – why did the journal not publish more, why did the journal not make more revenue? Thus that line:

“their success will depend on accurate communication with readers”

Needs a whole lot more explanation and definition for me to able to understand it, let alone possibly agree it could be true. I can’t approve it until I understand exactly what the authors mean here. If it is not essential to the work, perhaps just strike it out? Talk about eLife’s future instead perhaps – future is more neutral and less measure/viewpoint dependent?

Author Response: Thank you for these comments. We agree with you and we’ve adjusted the opening paragraph to remove ambiguous/subjective statements “improvement” and “success”, avoid the implication that open peer review is a novel innovation, and clarify that our focus is on eLife’s new vocabulary:

“Peer review is usually a black box — readers only know that a research paper eventually surpassed some ill-defined threshold for publication and rarely see the more nuanced evaluations of the reviewers and editor (Vazire, 2021). A minority of journals challenge this convention by making peer review reports publicly available (Wolfram et al., 2020). One such journal, *eLife,* also accompanies articles with short evaluation statements ("*eLife* assessments") representing the consensus opinions of editors and peer reviewers (Eisen et al., 2022). *eLife* recently stated that these assessments will use phrases drawn from a common vocabulary (Table 1) to convey two evaluative dimensions: (1) "significance"; and (2) "strength of support" (for details see *eLife's New Model*, 2022). For example, a study may be described as having "landmark" significance and offering "exceptional" strength of support (for a complete example, see Box 1). The phrases are drawn from "widely-used expressions" in prior *eLife* assessments and the stated goal is to 'help convey the views of the editor and the reviewers in a clear and consistent manner' (*eLife's New Model*, 2022). Here we outline a study intended to assess whether the language used in *eLife* assessments is perceived clearly and consistently by potential readers. We also propose and assess alternative language that may improve communication."

34. Linguistic pedantry: "...used in eLife assessments is perceived as intended by potential readers"

    I think as both the assessments are plural and the readers are plural then the perception of those assessments by readers should also be plural? "is perceived" -> "are perceived" .

Author Response: Removed.

35. Have the authors considered whether eLife's common vocabulary might be _intentionally_ ambiguous and thus not always necessarily so transparently easy to order on a scale? I don't know if that is the case, but if it really were so easy to number the terms why wouldn't eLife just use a 0 to 10 number scale? Is ambiguity in assessment not 'romantic' or some such, a specific and deliberate avoidance of ruthless and inhuman clarity?

Author Response: Please see our response to comment 3.

36. On the suggested alternative vocabulary, have you thought about suggesting “no importance” ? Given the prevalence of AI-generated texts and fraud in the publishing system, I can clearly see times in which a reviewer will want & need for the sake of integrity to reach for “no importance” or “zero importance” to accurately represent the utter rubbish they have read. I say this even despite footnote 1. Yes, eLife will likely be able to filter out some of the trash before sending for peer review, but perhaps not all of it.

Author Response: We think our current set of phrases is sufficient for an initial empirical demonstration as it’s unclear if such a “no importance” categorization would be useful given the initial manuscript triage by editors.

37. I am certainly not an expert in the science of misperception. Far far from it – I have no research experience in this area. However I do think it is naïve to think that simple English language words & phrases will ever ever be necessarily interpreted exactly/identically the same by people from e.g. different cultures and different parts of the world. Language and meaning doesn’t always have absolute precision, especially across different cultural contexts.

The word ‘tabled’ is a classic & extreme example of this. A British person would probably think it is an idea on the agenda for discussion, whilst an American might think it has been postponed or cancelled – despite being the exact same word, and same spelling and written in “English”. It has two very different interpretations by two different communities of readers - both of whom are fluent in "English".

Author Response: Thank you for raising this. Note that we are not expecting that people will interpret words in ‘exactly’ the same way; only that some phrases are interpreted more consistently than others. Also note that our alternative vocabulary deliberately does *not* use different words to represent aspects of the same evaluative dimension. We propose using modifications of the same word — “very high importance”, “high importance”, “moderate importance”, etc.

38. Regarding aim Two

The authors themselves allude to this but what if there isn't necessarily a single unidimensional ranking for strength of support. What if instead there were three or four dimensions upon which a manuscript could be assessed. eLife seems to only band it into two: 'significance' and 'Strength of support' – perhaps this is where an issue might lie?

Author Response: We agree that these dimensions could theoretically be split into more granular dimensions; we don't think that necessarily undermines the pragmatic utility of reporting editor/reviewer opinions of a single 'parent' dimension.

39. Comments on the proposed methodology

As far as I know, eLife was not created or intended for _just_ the UK and USA demographic, neither on the author-side nor the reader-side. The authorship profile in this journal certainly is more diverse than just this, and is publicly available data (e.g. at Lens: https://www.lens.org/lens/search/scholar/analysis?p=0&n=10&s=date_published&d=%2B&f=false&e=false&l=en&authorField=author&dateFilterField=publishedYear&orderBy=%2Bdate_published&presentation=false&preview=true&stemmed=true&useAuthorId=false&sourceTitle.must=eLife)

As I have alluded to earlier, I think it is important to account for cultural/geographical variance in the interpretation of English-language words and phrases. Unless I missed it, I see nothing in the method that would prevent the participants from comprising 100% of e.g. people born and raised in the USA. Which would be useful in some respects (some data better than no data), but not a global view, and eLife is globally read and diversely authored-in (not all countries e.g. zero Zambia/Chad/Sudan-based institutional author affiliations so far) as far as I know.

The WEIRD (Western, educated, industrialized, rich and democratic) bias is well known in psychology and I’m surprised I haven’t seen this registered report do more to ameliorate it (okay, maybe the E is needed in this case given the high-level content, but the potential for WIRD bias still needs to be acknowledged and ameliorated).

I see the authors refer to it as a "convenience sample" in the limitations section, and I acknowledge that. But I think just a little bit more effort and a slightly more complex design would significantly improve the validity and robustness of this research.

Ideally, I would like to see the participant sample stratified either by where existing eLife readers are known to come from (if that data is available) OR by where English language scientific articles are produced from, for which data is most certainly available e.g. https://en.wikipedia.org/wiki/List_of_countries_by_number_of_scientific_and_technical_journal_articles The participant sample absolutely must comprise at least some people living in these major consumer&producer (of science) countries: China, US, UK, India, Germany, Italy, Japan, Canada, Russia, France, Australia, Spain, South Korea, Brazil, Iran, Netherlands, Turkey, Poland, Indonesia, Switzerland…

Given the stratification required and the chance that country/culture has an effect on perception of English-language words, I suspect the minimum sample size will also need to be enlarged to accommodate the ability to compare perceptions between countries, which clearly won’t be statistically meaningful to do if only 300 participants (all countries) are sampled.

It might also be interesting to consider, for readers whose first language is NOT English (whilst also being fluent in English), how exactly do they choose to read review reports at eLife ? A small proportion for instance may opt to machine-translate from the English report into their first language, and _then_ read the report. One must perhaps examine how the two suggested vocabularies are machine-translated into other languages and if any peculiarities arise from that. Facebook machine translation once translated ‘good morning’ into ‘attack them’ leading to a Palestinian man being arrested (see https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest). Language, translation, and perception is hard to get right.

Author Response: Thank you for these comments. We recognize that eLife likely has a global and varied readership, and that if our study sample does not represent that readership, this will reduce the generalizability of the results. The hard question is what we can do to address that. Ideally, we would collect a sample that is representative of the target population (i.e., potential readers of scientific articles); however, doing so would require that (a) we know which demographic characteristics are relevant; (b) we know the distribution of those demographic characteristics in the target population; (c) we have some practical means of obtaining a representative sample. Addressing these issues is either not possible or beyond our resources/scope (also see our response to comment 4). We also do not think it is feasible to evaluate the accuracy of machine-translation for multiple languages; there are too many unknowns (e.g., what translation tools readers are using) and extra resources would be required to e.g., hire translators to validate the machine translations.

40. Comments on the proposed analysis plan

The analysis plan needs to factor-in country/culture of the study participant and analyze potential differences of interpretation between source countries. Not controlling for this potential variance undermines the usefulness of the proposed design.

Author Response: See response to comment 39.