Finding the right words to evaluate research: An empirical appraisal of *eLife*’s assessment vocabulary

Tom E. Hardwicke1\* ([0000-0001-9485-4952](https://orcid.org/0000-0001-9485-4952)), Sarah Schiavone1 ([0000-0002-8829-137X](https://orcid.org/0000-0002-8829-137X)), Beth Clarke1 ([0000-0002-7355-6718](https://orcid.org/0000-0002-7355-6718)), Simine Vazire1 ([0000-0002-3933-9752](https://orcid.org/0000-0002-3933-9752))

1Melbourne School of Psychological Sciences, University of Melbourne, Australia

\*Corresponding author: Tom E. Hardwicke (tom.hardwicke@unimelb.edu.au)

**Competing interests statement:** All authors declare they have no conflicts of interest.

**Funding statement:** This research will be supported by funding from the Melbourne School of Psychological Sciences, University of Melbourne.

**Ethics statement:** This study was approved by a University of Melbourne ethics board (project ID: 26411).

**Open research statement:** All data, materials, and analysis scripts will be made publicly available on the Open Science Framework (<https://osf.io/jpgxe/>). A reproducible version of the manuscript and associated computational environment will be made available in a Code Ocean container.

# INTRODUCTION

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)](https://www.zotero.org/google-docs/?2aSgwC). A minority of journals challenge this convention by making peer review reports publicly available [(Wolfram et al., 2020)](https://www.zotero.org/google-docs/?D2VvXO). 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)](https://www.zotero.org/google-docs/?Bo3X3Q). *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](https://www.zotero.org/google-docs/?6ZQ9Xw) *[eLife’s New Model](https://www.zotero.org/google-docs/?6ZQ9Xw)*[, 2022)](https://www.zotero.org/google-docs/?6ZQ9Xw). 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’ (*e[Life’s New Model](https://www.zotero.org/google-docs/?4cDwkd)*[, 2022](https://www.zotero.org/google-docs/?4cDwkd)). 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.

Our understanding (based on *e*[*Life’s New Model*, 2022](https://www.zotero.org/google-docs/?4cDwkd)) is that *eLife* intends the common vocabulary to represent different degrees of each evaluative dimension on an ordinal scale (e.g., “landmark” findings are more significant than “fundamental” findings, and so forth); however, in our view the intended ordering is sometimes ambiguous or counterintuitive. For example, it does not seem obvious to us that an “important” study is necessarily more significant than a “valuable” study, nor does a “compelling” study seem necessarily stronger than a “convincing” study. Additionally, several phrases like “solid” and “useful”, could be broadly interpreted, leading to a mismatch between intended meaning and perceived meaning. The phrases also do not cover the full continuum of measurement and are unbalanced in terms of positive and negative phrases[[1]](#footnote-1). For example, the “significance” dimension has no negative phrases — the scale endpoints are “landmark” and “useful”. We also note that the definitions provided by *eLife* do not always map onto gradations of the same construct. For example, the *eLife* definitions of phrases on the significance dimension suggest that the difference between “useful”, “valuable”, and “important” is a matter of breadth/scope (whether the findings have implications beyond a specific subfield) whereas the difference between “fundamental” and “landmark” is a matter of degree. In short, we are concerned that several aspects of the *eLife* vocabulary may undermine communication of research evaluations to readers.

In Table 2, we outline an alternative vocabulary that is intended to overcome the potential issues with the *eLife* vocabulary. Phrases in the alternative vocabulary explicitly state the relevant evaluative dimension (e.g., “support”) along with a modifying adjective that unambiguously represents degree[[2]](#footnote-2) (e.g., “very low”). The alternative vocabulary is also intended to cover the full continuum of measurement and be balanced in terms of positive and negative phrases. We have also renamed “significance” to “importance” to avoid any confusion with statistical significance. We hope that these features will facilitate alignment of readers’ interpretations with the intended interpretations, improving the efficiency and accuracy of communication.

**Table 1**. Phrases (italicised) from the *eLife* vocabulary representing two evaluative dimensions: significance and strength of support. The significance dimension is represented by five phrases and the strength of support dimension is represented by six phrases. In a particular *eLife* assessment, readers only see one phrase from each of the evaluative dimensions. Phrases are accompanied by *eLife* definitions, but these are not shown in *eLife* assessments (though some words from the definitions may be used).

|  |  |
| --- | --- |
| ***eLife* vocabulary** | |
| **Significance** | **Strength of support** |
| *Landmark*: findings with profound implications that are expected to have widespread influence | *Exceptional*: exemplary use of existing approaches that establish new standards for a field |
| *Fundamental*: findings that substantially advance our understanding of major research questions | *Compelling*: evidence that features methods, data and analyses more rigorous than the current state-of-the-art |
| *Important*: findings that have theoretical or practical implications beyond a single subfield | *Convincing*: appropriate and validated methodology in line with current state-of-the-art |
| *Valuable*: findings that have theoretical or practical implications for a subfield | *Solid*: methods, data and analyses broadly support the claims with only minor weaknesses |
| *Useful*: findings that have focused importance and scope | *Incomplete*: main claims are only partially supported |
|  | *Inadequate*: methods, data and analyses do not support the primary claims |

**Box 1**. A complete example of a recent *eLife* assessment. This particular example uses the phrase “important”, to convey the study’s significance, and the phrase “compelling”, to convey the study’s strength of support.

|  |
| --- |
| “The overarching question of the manuscript is important and the findings inform the patterns and mechanisms of phage-mediated bacterial competition, with implications for microbial evolution and antimicrobial resistance. The strength of the evidence in the manuscript is compelling, with a huge amount of data and very interesting observations. The conclusions are well supported by the data. This manuscript provides a new co-evolutionary perspective on competition between lysogenic and phage-susceptible bacteria, that will inform new studies and sharpen our understanding of phage-mediated bacterial co-evolution.” [(Rendueles et al., 2023)](https://www.zotero.org/google-docs/?IAnHj0). |

**Table 2**. Phrases (italicised) from the alternative vocabulary representing two evaluative dimensions: importance and strength of support. Each dimension is represented by five phrases.

|  |  |
| --- | --- |
| **Alternative vocabulary** | |
| **Importance** | **Strength of support** |
| *Very high importance* | *Very strong support* |
| *High importance* | *Strong support* |
| *Moderate importance* | *Moderate support* |
| *Low importance* | *Weak support* |
| *Very low importance* | *Very weak support* |

The utility of *eLife* assessments will depend (in part) on whether readers interpret the common vocabulary in the manner that *eLife* intends. Mismatches between *eLife*’s intentions and readers’ perceptions could lead to inefficient or inaccurate communication. 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. Our goal is not to disparage *eLife’s* progressive efforts, but to make a constructive contribution towards a more transparent and informative peer review process. We hope that a vocabulary with good empirical performance will be more attractive and useful to other journals considering adopting *eLife*’s approach.

Our study is modelled on prior studies that report considerable individual differences in people’s interpretation of probabilistic phrases [(Budescu et al., 2014; Budescu & Wallsten, 1985; Lichtenstein & Newman, 1967; Reagan et al., 1989; Theil, 2002; Wallsten et al., 1986; Willems et al., 2020)](https://www.zotero.org/google-docs/?zN8yJq). In a prototypical study of this kind, participants are shown a probabilistic statement like “It will probably rain tomorrow” and asked to indicate the likelihood of rain on a scale from 0-100%. Analogously, we intend to show people statements about hypothetical scientific studies using phrases drawn from the *eLife* vocabulary or the alternative vocabulary, and ask them to indicate the study’s significance/importance or strength of support on a scale from 0-100. These responses will allow us to gauge the extent to which people’s interpretations of the vocabulary phrases are consistent with each other and consistent with the intended rank order.

### Research aims

Our overarching goal is to identify language for conveying evaluations of scientific papers in a way that is clear and useful to readers. We hope that this will make it easier for other journals/platforms to follow in *eLife*'s footsteps and move towards more transparent and informative peer review.

With this overall goal in mind, we have three specific research aims:

* Aim One. To what extent do people share similar interpretations of phrases used to describe scientific research?
* Aim Two. To what extent do people’s (implicit) ranking of phrases used to describe scientific research align with (a) each other; and (b) with the intended ranking?
* Aim Three. To what extent do different phrases used to describe scientific research elicit overlapping interpretations and do those interpretations imply broad coverage of the underlying measurement scale?

# METHODS

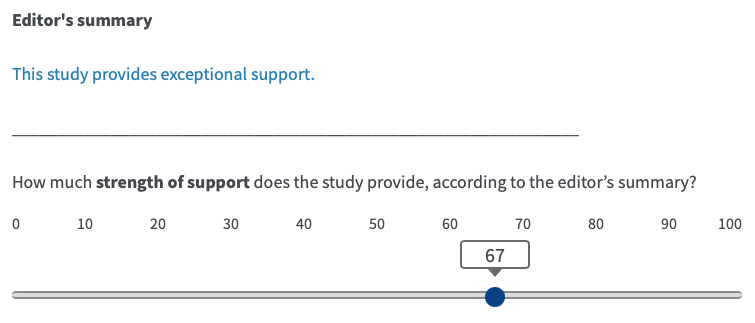
### Design

This is an experiment with a repeated-measures design. Participants will be shown short statements that describe hypothetical scientific studies in terms of their significance/importance or strength of support using phrases drawn from the *eLife* vocabulary (Table 1) and from the alternative vocabulary (Table 2). The statements are organised into four blocks based on vocabulary and evaluative dimension; specifically, block one: *eLife-significance* (5 statements), block two: *eLife-support* (6 statements), block three: *alternative-importance* (5 statements), block four: *alternative-support* (5 statements). Each participant will see all 21 phrases and respond using a 0-100% scale to indicate their belief about each study’s significance/importance or strength of support.

### Materials

There are 21 statements that describe hypothetical scientific studies using one of the 21 phrases included in the two vocabularies (Table 1). Statements refer either to a study’s strength of support (e.g., Figure 1) or a study’s significance/importance (e.g., Supplementary Figure A1). For the alternative vocabulary, we use the term “importance” rather than “significance”. To ensure the statements are grammatically accurate, it is necessary to use slightly different phrasing when communicating significance with the *eLife* vocabulary (“This is an [phrase] study”) compared to communicating importance with the alternative vocabulary (“This study has [phrase] importance”; e.g., Supplementary Figure A2). Additionally, there is one attention check statement (Supplementary Figure A3), a question asking participants to confirm their highest completed education level (options: Undergraduate degree (BA/BSc/other)/Graduate degree (MA/MSc/MPhil/other)/Doctorate degree (PhD/other)/Other), and a question asking participants the broad subject area of their highest completed education level (options: Arts & Humanities/Life Sciences and Biomedicine/Physical Sciences/Social Sciences/Other). The veridical materials are available at https://osf.io/jpgxe/

**Figure 1**. An example summary statement referring to a study’s strength of support and the corresponding response scale with an arbitrary response shown.



### Sample

*Sample source.* Participants will be recruited from the online participant recruitment platform Prolific Academic (<https://www.prolific.co/>). 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.

*Sample size.* Our target sample size is 300 participants (without exclusions).

*Sample size justification.* The target sample size is based on our resource constraints and should yield reasonable statistical power and precision (see Supplementary Information B for details).

*Inclusion criteria*. 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).

### Procedure

1. After responding to the study advert (<https://osf.io/a25vq>), participants will read an information sheet (<https://osf.io/39vay>), and provide consent (<https://osf.io/xdar7>). During this process, they will be told that the study seeks to understand “how people perceive words used to describe scientific studies so we can improve communication of research to the general public.”
2. Participants will complete the task remotely online via the Qualtrics platform. Before starting the main task, they will read a set of instructions and respond to a practice statement (Supplementary Information C).
3. For the main task, statements will be presented sequentially and participants will respond to them in their own time. The order of presentation will be randomized, both between and within the four blocks of statements. 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. A single ‘attention check’ statement (Supplementary Figure A3) will appear after all four blocks have been completed.
4. Participants must respond to each statement before continuing to the next statement. The response slider can be readjusted as desired until the ‘next’ button is pressed, after which participants cannot return to prior statements or edit prior responses.
5. After responding to all 21 statements and the attention check, participants will be shown a debriefing document (<https://osf.io/a9gve>).

### Analysis plan

Our analysis plan is informed by data simulations and mock analyses reported in a dynamic document written in Quarto (<https://osf.io/bc6vk>). A HTML version of the document is available (<https://osf.io/8v9n5>) which can be downloaded and opened in any web browser. The document walks through the planned analyses step-by-step, performs the analyses on simulated data, and illustrates the (simulated) output of the analyses.

*Exclusion criteria.* Before proceeding with the substantive data analysis, 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[[3]](#footnote-3); (2) they do not respond to all 21 statements; or (3) they fail the attention check. The number of exclusions and reason for exclusion will be reported.

*Analyses for Research Aim 1 (*To what extent do people share similar interpretations of phrases used to describe scientific research?). For continuous responses to each phrase, we will report medians, 25th and 75th percentiles, and the interquartile range. For all of these estimates we will report 95% confidence intervals, bootstrapped with the percentile method [(Rousselet et al., 2021)](https://www.zotero.org/google-docs/?IWDSyX). We will visualize the data with kernel density distributions, which represent the relative probability of responses (akin to a ‘smoothed histogram’; [Wilke, 2019)](https://www.zotero.org/google-docs/?h7LC7l).

*Analyses for Research Aim 2 (*To what extent do people’s (implicit) ranking of phrases used to describe scientific research align with (a) each other; and (b) with the intended ranking?). For each participant, continuous responses will be converted to rankings (‘observed ranking’) for each evaluative dimension of each vocabulary. If a participant provides the same response to more than one phrase, the relevant ranks will be randomly allocated among the affected phrases. We will compute the proportion of participants who ranked the words in particular orders (‘ranking sequences’), with particular emphasis on identifying how many observed rankings aligned with the intended rankings for each vocabulary. Multinomial proportions will be accompanied by 95% confidence intervals computed with the Sison-Glaz method [(Sison & Glaz, 1995)](https://www.zotero.org/google-docs/?2HpKI8).

To compare ranking accuracy between the two vocabularies, we will collapse the observed rankings that did not match the intended ranking into a single ‘No Match’ category, and use a McNemar test to examine whether a match was more likely for the alternative vocabulary relative to the eLife vocabulary. Specifically, we will test the null hypothesis that the number of participants who responded correctly to the *eLife* vocabulary and incorrectly to the alternative vocabulary was equal to the number of participants who responded correctly to the alternative vocabulary and incorrectly to the eLife vocabulary. 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. We will perform an ‘exact’ McNemar test (alpha = .05) and report the odds ratio with Clopper Pearson 95% confidence intervals, adjusted with the ‘midp’ method, as recommended by [Fagerland et al. (2013)](https://www.zotero.org/google-docs/?PGOlR5). These analyses will be performed separately for each evaluative dimension.

The above approach does not take into account the fact that, among the “no match” rankings, some ranking sequences will be more similar to the intended ranking sequence than others. One method to quantify the similarity between rankings is to compute Kendall’s tau distance (*Kd*) — a metric that describes the difference between two lists in terms of the number of adjacent pairwise swaps required to convert one list into the other [(Kendall, 1938; van Doorn et al., 2021)](https://www.zotero.org/google-docs/?B9VRuW). A larger *Kd* represents greater dissimilarity between two ranking sequences. Will we compute the *Kd* for observed rankings relative to the corresponding intended ranking and present the distribution of *Kd* scores for each evaluative dimension of each vocabulary. We will report the normalized *Kd* because one vocabulary set has six phrases and the other sets have five.

Finally, in order to identify which specific phrases (if any) tend to be misranked, we will plot heatmaps showing the proportion of concordant and discordant rankings.

*Analyses for Research Aim 3 (*To what extent do different phrases used to describe scientific research elicit overlapping interpretations and do those interpretations imply broad coverage of the underlying measurement scale?). We have not planned any formal analyses for Research Aim 3. We intend to address this question informally through visual examination of the kernel density distributions plotted for each phrase.

# DISCUSSION & LIMITATIONS

Below we outline several notable study limitations and points of discussion that we anticipate informing our interpretation of the results. We plan to integrate and reformat these points in our Stage Two report, but include them here to clarify the goals and limitations of our project.

* Our study does not address whether editor/reviewer opinions provide (a) valid assessments of studies; or (b) whether the vocabularies provide valid measurements of those opinions. We also note that *eLife* assessments are formed via consensus, rather than representing the opinions of an individual, which raises questions about how social dynamics affect the evaluation outcomes. It may be more informative to solicit and report individual assessments from each peer reviewer and editor, rather than force a consensus. Although these are important issues, they are beyond the scope of this study. This study is focused on clarity of communication, i.e., the extent to which readers’ interpretations of the vocabularies are aligned with their intended interpretation.
* 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.
* To maintain experimental control, we intend to present participants with very short statements that differ only in terms of the phrases we wish to evaluate. In practice however, the phrases will be embedded in a paragraph of text (Box 1) which may also contain “aspects” of the vocabulary definitions (Table 1) “when appropriate” [(*eLife’s New Model*, 2022)](https://www.zotero.org/google-docs/?nOZPhI). Future empirical assessments should explore the impact of these factors. For example, it will be interesting to see whether participants are even aware of the relevant phrases and the dimensions they are intended to represent when they appear in the context of a paragraph summary. Future improvements might explore combining *eLife* assessments with a visualisation of the full measurement continuum, as has been examined in the context of communicating probabilistic statements related to climate change [(Budescu et al., 2014)](https://www.zotero.org/google-docs/?ucFsP6). Such visualisations may be especially beneficial for non-native English speakers as they should theoretically help to disambiguate the vocabularies.
* In our study, participants will be asked to respond to phrases with a point estimate, whereas arguably they interpret phrases in terms of a range of plausible values [(Reagan et al., 1989; Wallsten et al., 1986)](https://www.zotero.org/google-docs/?o722yp). However, asking participants to respond with a range, rather than a point estimate, creates technical and practical challenges and complicates data analysis. We therefore decided to focus on point estimates, at least for a first empirical exploration of this issue.
* Finally, we note that currently *eLife* assessments are performed at the article level, which may be inappropriate when an article contains multiple scientific claims (because the evaluation of importance and strength of support may vary between claims).

# REFERENCES

[Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508–512. https://doi.org/10.1038/nclimate2194](https://www.zotero.org/google-docs/?t4LTbI)

[Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X](https://www.zotero.org/google-docs/?t4LTbI)

[Eisen, M. B., Akhmanova, A., Behrens, T. E., Diedrichsen, J., Harper, D. M., Iordanova, M. D., Weigel, D., & Zaidi, M. (2022). Peer review without gatekeeping [editorial]. *eLife*, *11*, e83889. https://doi.org/10.7554/eLife.83889](https://www.zotero.org/google-docs/?t4LTbI)

[*eLife review process FAQs*. (2023, May 2). https://elife-cdn.s3.amazonaws.com/documents/eLife\_review\_process\_FAQs.pdf](https://www.zotero.org/google-docs/?t4LTbI)

[*eLife’s New Model: What is an eLife assessment?* (2022, October 20). eLife; eLife Sciences Publications Limited. https://elifesciences.org/inside-elife/db24dd46/elife-s-new-model-what-is-an-elife-assessment](https://www.zotero.org/google-docs/?t4LTbI)

[Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, *13*(1), 91. https://doi.org/10.1186/1471-2288-13-91](https://www.zotero.org/google-docs/?t4LTbI)

[Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81](https://www.zotero.org/google-docs/?t4LTbI)

[Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, *9*, 563–564.](https://www.zotero.org/google-docs/?t4LTbI)

[Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*(3), 433–442. https://doi.org/10.1037/0021-9010.74.3.433](https://www.zotero.org/google-docs/?t4LTbI)

[Rendueles, O., de Sousa, J. A., & Rocha, E. P. (2023). Competition between lysogenic and sensitive bacteria is determined by the fitness costs of the different emerging phage-resistance strategies. *eLife*, *12*, e83479. https://doi.org/10.7554/eLife.83479](https://www.zotero.org/google-docs/?t4LTbI)

[Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920911881. https://doi.org/10.1177/2515245920911881](https://www.zotero.org/google-docs/?t4LTbI)

[Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, *90*(429), 366–369. https://doi.org/10.1080/01621459.1995.10476521](https://www.zotero.org/google-docs/?t4LTbI)

[Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: A meta-analysis. *Journal of Risk Research*, *5*(2), 177–186. https://doi.org/10.1080/13669870110038179](https://www.zotero.org/google-docs/?t4LTbI)

[van Doorn, J., Westfall, H. A., & Lee, M. D. (2021). Using the weighted Kendall Distance to analyze rank data in psychology. *The Quantitative Methods for Psychology*, *17*(2), 154–165. https://doi.org/10.20982/tqmp.17.2.p154](https://www.zotero.org/google-docs/?t4LTbI)

[Vazire, S. (2021). Opening the Black Box of Peer Review. *Physics*, *14*, 169. https://physics.aps.org/articles/v14/169](https://www.zotero.org/google-docs/?t4LTbI)

[Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348–365. https://doi.org/10.1037/0096-3445.115.4.348](https://www.zotero.org/google-docs/?t4LTbI)

[Wilke, C. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures* (First edition). O’Reilly Media.](https://www.zotero.org/google-docs/?t4LTbI)

[Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles—A risk for miscommunication. *Journal of Science Communication*, *19*(02), A03. https://doi.org/10.22323/2.19020203](https://www.zotero.org/google-docs/?t4LTbI)

[Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: Promoting transparency in open science. *Scientometrics*, *125*(2), 1033–1051. https://doi.org/10.1007/s11192-020-03488-4](https://www.zotero.org/google-docs/?t4LTbI)

# 

# SUPPLEMENTARY INFORMATION

### SUPPLEMENTARY INFORMATION A: Supplementary materials

**Supplementary Figure A1**. An example summary statement referring to a study’s significance and the corresponding response scale with an arbitrary response shown.

A screenshot of a survey

Description automatically generated

**Supplementary Figure A2**. An example summary statement referring to a study’s importance and the corresponding response scale with an arbitrary response shown.

**A screenshot of a survey

Description automatically generated**

**Supplementary Figure A3**. Attention check statement.

A screenshot of a computer screen

Description automatically generated

### SUPPLEMENTARY INFORMATION B: Sample size planning

We firstly decided that 300 participants was a reasonable sample size target given our resources. We then evaluated the expected statistical power and precision in a plausible scenario, assuming a sample size of 300 and a two-sided test with alpha = .05. For shorthand, we refer to a ‘correct response’ where the observed ranking matches the intended ranking.

In a scenario where 10% of people respond correctly to the eLife vocabulary and incorrectly to the alternative vocabulary, and 20% of people respond correctly to the alternative vocabulary and incorrectly to the *eLife* vocabulary, this would yield a McNemar odds ratio of 2, 95% confidence intervals [1.3-3.1] and statistical power of 0.87 with an exact McNemar test. Analysis code documenting these calculations is available at <https://osf.io/8v9n5> (under the heading “Sample size planning”).

### 

### SUPPLEMENTARY INFORMATION C: Task instructions

Before starting the study, participants will be presented with the instructions shown in Supplementary Box C1 and have the opportunity to respond to a practice statement. At the start of each block they will be shown the instructions shown in Supplementary Box C2.

**Supplementary Box C1**. Pre-study instructions.

|  |
| --- |
| PAGE 1  Thank you for your participation in this study. Please ensure you are in a quiet, distraction-free environment before starting the task. Please give the task your full attention, it will only take about 10 minutes of your time.  PAGE 2  Imagine that you visit the website of a scientific journal to read some articles reporting scientific studies. You see that each article is accompanied by a short summary statement expressing the editor’s opinion of the study report in the article.  We are going to show you 22 statements describing the editor’s opinion of 22 different research studies. For each statement, we’d like you to tell us what you think about a particular aspect of the study using a slider on a scale ranging from 0 to 100%.  After each statement, you will complete a 15 second task involving simple multiplication questions.  PAGE 3  Here’s an example before we start. The editor’s summary appears in blue. In this example, your task is to rate how **clearly written** you think the article is based on the editor’s summary statement. You can click and drag the slider to choose your response.  Practice using the slider now and then click next when you are ready to start the study. If you are having technical problems or anything is unclear, please contact tom.hardwicke@unimelb.edu.au |

A screenshot of a computer

Description automatically generated

**Supplementary Box C2**. Pre-block instructions.

|  |
| --- |
| STRENGTH OF SUPPORT BLOCKS  You will now see statements about the **strength of support** offered by 5/6 different research studies. These statements represent the journal editor’s opinion about the **strength of support** each study offers towards its main claims.  SIGNIFICANCE BLOCK  You will now see statements about the **significance** of 5 different research studies. These statements represent the journal editor’s opinion about the **significance** of each of the study’s main claims.  IMPORTANCE BLOCK  You will now see statements about the **importance** of 5 different research studies. These statements represent the journal editor’s opinion about the **importance** of each of the study’s main claims. |

**SUPPLEMENTARY INFORMATION D. Peer Community in Registered Reports Design Template.**

**Supplementary Table D1.** Peer Community in Registered Reports Design Template.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question** | **Hypothesis** | **Sampling plan** | **Analysis Plan** | **Rationale for test sensitivity** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| Research Aim 1: To what extent do people share similar interpretations of phrases used to describe scientific research? | N/A | See Main Text — sub-section “Sample” and Supplementary Information B. | See Main Text — *Analyses for Research Aim 1.* | N/A. No hypotheses. | Descriptive. | N/A (applied research question) |
| Research Aim 2: To what extent do people’s (implicit) ranking of phrases used to describe scientific research align with (a) each other; and (b) with the intended ranking? | N/A for (a). For (b), we hypothesize that the alternative vocabulary will have higher ranking accuracy than the eLife vocabularly. | See Main Text — sub-section “Sample” and Supplementary Information B. | See Main Text — *Analyses for Research Aim 2.* | See Supplementary Information B | Mostly descriptive. For the hypothesis test, a significant result will indicate that ranking accuracy is higher for one vocabulary relative to the other, and it would be surprising to observe this difference if the null hypothesis were correct. A non-significant result will imply that the observed results are not surprising under the null hypothesis, and we therefore do have grounds to conclude that the ranking accuracy differs between the two vocabularies. | N/A (applied research question) |
| Research Aim 3: To what extent do different phrases used to describe scientific research elicit overlapping interpretations and do those interpretations imply broad coverage of the underlying measurement scale? | N/A | See Main Text — sub-section “Sample” and Supplementary Information B. | See Main Text — *Analyses for Research Aim 3.* | N/A. No hypotheses. | Descriptive. | N/A (applied research question) |

1. Note that *eLife* does not intend to send all submitted manuscripts for peer review [(*eLife Review Process FAQs*, 2023)](https://www.zotero.org/google-docs/?SFf0kM), so it is possible they have omitted negative phrases on the grounds that less-than-useful papers should not make it through initial editorial triage. Nevertheless, we think it would be prudent to include negative phrases in the vocabulary because peer reviewers may disagree with the editor, or the editor may change their mind. [↑](#footnote-ref-1)
2. We suggest that if *eLife* wants to convey breadth/scope as well as degree, this should be represented by a separate dimension. [↑](#footnote-ref-2)
3. Based on piloting, we estimate that a reasonable person focused solely on the task should take approximately 10 to 15 minutes to complete it. [↑](#footnote-ref-3)