Title: Development and application of a comprehensive glossary for the identification of statistical and methodological concepts in peer review reports

Authors: Ivan Buljan,[a,b]* Daniel Garcia-Costa,[c] Francisco Grimaldo,[c] Richard A. Klein,[d] Marjan Bakker,[d] Ana Marušić[a]

Affiliations:
[a] Department of Research in Biomedicine and Health, Center for Evidence-based Medicine, University of Split School of Medicine, Split, Croatia, ibuljan@mefst.hr , amarusic@mefst.hr
[b] Department of Psychology, Faculty of Humanities and Social Sciences, University of Split, Split, Croatia, ibuljan@ffst.hr
[c]Department of Computer Science, University of Valencia, Burjassot, Spain, francisco.grimaldo@uv.es, daniel.garcia@uv.es
[d]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands, R.A.Klein@tilburguniversity.edu, M.Bakker_1@tilburguniversity.edu

*Corresponding author.
Address for correspondence:
Assist. Prof. Ivan Buljan, PhD
Faculty of Humanities and Social Sciences
Poljicka 35
21000 Split, Croatia
ivan.buljan@ffst.hr

Highlights

- We constructed a glossary with terms related to research methodology and statistics
- The glossary was used to assess issues in peer review reports.
- The glossary identified statistical and methodological issues in the reports.
- Reviewer rejection was related to methodological but not statistical issues.
- Statistical concepts were more present in research studies.

Abstract:

The assessment of problems identified by peer researchers during peer review is difficult because the content of these reports is typically confidential. The current study sought to construct and apply a glossary for the identification of methodological and statistical concepts mentioned in peer review reports. Three assessors created a list of 1,036 different terms in 19 categories. The glossary was tested on the confidential PEERE database, a sample of 496,928 peer review reports from various scientific disciplines. The most frequently mentioned terms were related to data presentation (found in 40.3% of the reports) and parametric descriptive statistics (33.3%). Review reports suggesting a rejection were more likely to mention methodological issues, whereas statistical issues were raised more frequently in review reports recommending revisions. Across disciplines, methodological issues were more frequently mentioned in social sciences (64.1%), while health and medical sciences were more predictive for the identification of statistical issues (40.1%). Female reviewers identified more statistical issues compared to male reviewers. These results indicate that the glossary could be used as an additional tool for the assessment of the content of peer review reports and for understanding what help authors may need in writing research articles.

Keywords (6):

Peer review; dictionary-based approach; statistical glossary; discipline differences; PEERE database; research quality assessment

1. Introduction

Peer review is a critical step in the assessment of research articles for publication. Although sometimes criticized for being flawed and biased (e.g. Huber et al., 2022; Sun et al., 2023), it still represents the "gold standard" in research quality control. Recent studies emphasize that peer reviewer's experience and greater knowledge of the topic can be critical in distinguishing between poor and good research (Zheng et al., 2023). Deeper insight into the content of peer review reports is critical to improving both the process of peer review, as well as informing about areas to target to improve the reporting of research results (Garcia-Costa et al., 2022a). Indeed, such information would be beneficial to all stakeholders in the publishing process, including authors, reviewers, publishers, and organizations that fund and perform research. For example, it could indicate the most common topics on which reviewers need to be educated. Due to the ever-present difficulty in attracting enough researchers to perform peer reviews (Fox, 2017), ensuring high quality among reviewers available for this work, and training new ones to handle the burden, is critical to keep the system sustainable.

In our approach, we limit our assessment to only quantitative studies. Building on the findings of our previous study (Garcia-Costa et al., 2022b), we conducted a study that would assess the content of peer reviews about quantitative research methods and statistics. In this study, due to the nature of peer review feedback, we assumed that the occurrence of specific words in a peer review reflected issues related to those words being critiqued by reviewers. Our dictionary-based approach therefore focused on counting words in the text, as a proxy measure for concepts, as used in previous studies (Reveilhac & Moreselli, 2022; Jacobucci, Ammerman & Wilcox, 2021; van Atteveldt, van der Velden & Boukes, 2021, Reagan et al., 2017; Meng, 2023).

2. Literature review

2.1. Availability of peer reviews

To reach valid conclusions, scientific studies must be appropriately designed. An important task for reviewers is to assess whether the design of a study was suitable for its aim, and suggest possible improvements when they identify weaknesses in the method. The content of peer review reports in the past was difficult to assess because peer review reports are often confidential, and rarely available for assessment on a larger scale (Squazzoni et al., 2020). In recent years, there has been an increase in the availability of peer review reports in databases like the PEERE database (Squazzoni et al., 2020) and peer review report characteristics like Elsevier's Peer Review Workbench (Petchiappan et al., 2022), which provide researchers with opportunities to explore the factors associated with peer review characteristics and patterns. Some journals (e.g., F1000 Research), decided to support transparent peer review by having post-publication review, while others have decided to opt for a consultation approach, where the peer reviewers get together and discuss the article in a panel (Urban et al., 2022). This increased availability of peer reviews allows for identifying characteristics of scientific gatekeeping and potential tools for improvement. At the same time, there is a growing need for the development of tools that will help to systematically analyze the vast amounts of very specific and very technical texts like peer review reports. Large language models (LLMs) like Chat GPT show a promising direction with automatized processes, but their use is still far from perfect and with still uncleared ethical questions (Lauer, Constant & Wernimont, 2023).

2.2. Assessment of peer review quality

Assessing the content of peer review reports and identifying the issues that reviewers frequently address can be done using a qualitative approach (Sizo et al., 2019). That approach would allow the identification of potential reasons for the acceptance or rejection of an article and give information about what reviewers consider quality research. One of the surrogate approaches is to study comments on preprints. Recent studies of reviewer comments on preprints offer evidence that reviewers mostly comment on issues related to methodology, literature and theoretical sources, and language use (Dimity, 2022). However, a qualitative approach would be limited to a small number of articles, preventing generalization. The second approach, the adoption of machine learning and artificial intelligence for such complex issues, is still in its infancy but is also a promising direction for future research (Checco et al., 2021; Chubb, Cowling & Reed, 2021; Horbach et al., 2022; Ghosal et al., 2022; Sun, 2024; Zheng et al., 2023). Studying the characteristics of the reviewers without the assessment of the content of the reports would be another approach. However, it would probably result in a biased assessment because it is not

106 focused on the core product (the report), with a potential outcome that reviewers with specific
107 characteristics are biased in their reports (e.g., Shopovski et al., 2020).

108 **2.3. Assessing methodology and statistics in peer review reports**

109 Studying how complex concepts related to methodology and statistics are assessed during peer
110 review on a larger scale is difficult without a large sample of peer review reports and a strong
111 methodology to identify those concepts. There is no commonly accepted approach in the assessment of
112 methodological and statistical issues in peer review reports, and there are several potential options. The
113 study using text mining approaches to identify trends in the analysis of openly available published articles
114 employed by different research studies showed that inferential statistical tests were the most prevalent,
115 but were slowly losing ground to machine learning and neural network approaches (Bolt et al., 2021).
116 Also, while some methodological aspects are extremely prevalent, others are rare and therefore present a
117 problem in training more data-hungry machine learning models (Kilicoglu et al., 2021; Thelwall et al.,
118 2023; Bharti et al, 2022). Finally, a mixed methods study on a smaller sample found that there were
119 reporting issues related to different aspects of methodology in quantitative studies, with descriptive
120 statistics parameters most commonly reported, and mention of existing outliers reported the least
121 frequently (Weiss et al., 2023). A potential alternative is the linguistic analysis of peer review reports,
122 focusing on methodological and statistical terms. In a recent study, a conclusion was that reviews from
123 social sciences are longer compared to the medical field, but the study was done on a relatively small
124 sample and with a glossary not specific to the research context (Perković-Palos et al., 2023). Currently,
125 there are glossaries for methodology and statistics developed specifically to assess peer review reports,
126 but those are still in need of refinement and expansion (Garcia-Costa et al, 2022a; Garcia -Costa et al,
127 2022b).

128 To assess the prevalence of methodological and statistical issues addressed in peer review reports,
129 our study consisted of two steps. In the first step, due to the lack of large glossaries related to statistics
130 and methodology, we had to construct a new glossary with words categorized in logical dimensions, so
131 that we could more precisely assess which methodological and statistical terms occur in peer reviews.

132 *Research aim 1: To construct and validate a glossary of terms related to methodological and*
133 *statistical aspects used in reviews by using a dictionary-based approach.*

134 Next, we applied this glossary to the PEERE dataset to identify how often these terms are used to
135 address methodological and statistical issues, and which were the most prominent issues. Besides the
136 identification of issues, we decided to answer practical questions: which review characteristics (e.g.,

reviewer gender or scientific discipline) are related to addressing methodological and statistical issues, and which issues are predictive of the final review recommendation?

Research aim 2: Applying the glossary to assess the prevalence and predictors of methodological and statistical issues addressed in the peer review report from the PEERE database.

The insights obtained from this research could serve as advice to the editors, peer reviewers, and authors in the assessment of research quality and identify targets for training and journal policy.

3. Material and Methods

3.1. Study design

This was a two-step study, in which we first created and validated the assessment tool (glossary), and secondly, applied it to the PEERE database to gain insights about common issues related to statistics and methodology raised in peer reviews. In this second step, we conducted a cross-sectional study by applying the newly constructed glossary to a large database of peer-review reports and the result was the number of terms related to various methodological and statistical categories, which was used as a proxy of issues addressed in reports (Squazzoni et al., 2020). The study was pre-registered on the Open Science Framework, and it is publicly available at: https://osf.io/tzb6n.

3.2. Construction of the glossary

In the first part of the study, we constructed a glossary with terms related to statistical and methodological concepts in research. The terms included in the glossary were identified through several sources: The Glossary of Statistical Terms (Stark, 2021), The Framework for Open and Reproducible Research Training Glossary (Parsons et al., 2022), and The Pocket Glossary for Commonly Used Research Terms (Hosko & Thyer, 2011). We classified the terms into several sub-categories, based on reporting guidelines for methodology and different types of statistical analysis (Simera et al, 2010). The categories falling under Methodology were: “Hypothesis”, “Study design”, “Sampling”, “Procedures”, “Biases”, and “Reporting guidelines”. The categories falling under Statistics were: “Distribution and probability”, “Meta-analysis”, “Variable characteristics and procedures”, “Accuracy and precision”, “Parametric descriptive parameters”, “Non-parametric descriptive statistics”, “Two group comparison statistics”, “Multi-group comparison”, “Diagnostics” (this section refers to the diagnosis of statistical models and tests), “Association”, “Effects”, “Data presentation”, and “Transparency”.

Three authors (IB, MB, and RK) reviewed the sources to identify all relevant terms for the glossary and classified them under the categories using their judgment. All three authors reviewed the classifications, discussed disputed terms in cases of disagreement, and came to a consensus. The categorization was presented to other members of the team which agreed with terms categorization. In the second part of the study, we expanded the glossary using a semi-automatic glossary-building approach, which ensures similar results to manually built dictionaries (Deng, 2019; Mpouli, 2020). This procedure followed five steps: 1) corpus creation, 2) pre-processing and cleaning 3) vector representation of the corpus, 4) term extraction, and 5) validation. In step 1, we collected and anonymized the text of review reports from our dataset. In step 2, we converted the text into lowercase, removed non-alphanumeric characters, trimmed white spaces and line breaks, tokenized web links and citations, and removed stop

words. In step 3, we built an unsupervised Word2Vec model using the H2O API (H2O.ai, 2022) in R (R Core Team, 2021) to create a vector representation of our corpus. In step 4, we used the Word2Vec model to search for near terms in all review texts. We extracted new terms by running the method 'findSynonyms' from the H2O API and selected the most frequent similar ones (i.e., those with a normalized score higher than 0.75) and listed them among list candidates. We set the threshold at 0.75 experimentally, since it was a conservative value that allowed us to extract a moderate number of relevant new terms and that ensured quick convergence in a reduced number of iterations. More concretely, steps 4 (term extraction) and 5 (validation) were repeated three times. The amount of new terms extracted in each iteration was about 600 and manual validation kept about 88% of the proposed terms. In step 5, the same three authors (IB, MB, and RK) manually validated the list of new terms by considering usage in different contexts and deciding whether to drop or retain each new term. This allowed us to use the output list of terms from the previous iteration as the input for the next iteration. Steps 4 and 5 were repeated until all new terms had low-frequency values. The procedure followed to develop our glossary does use NLP techniques (e.g. Word2Vec) to allow for complete, quick, and generalizable dictionary building and it is similar to previously used techniques (e.g. Han et al., 2022) While they focus on opinion words, we depart from a core of already accepted statistical terms to then enrich our glossary with the support of automatic up-to-date term extraction techniques and manual expert validation. The result of this procedure was a glossary containing 1,036 terms (available at https://osf.io/d34b9/).

3.3. Database and variables

The PEERE dataset is the result of a collaboration between publishers and researchers (Squazzoni et al., 2020). It contains over half a million peer review reports and related information from journals published by Elsevier. While we admit that there are other examples of available peer review reports, (e.g. PLoS database or the OpenReview website), the PEERE database is the only one that includes reviews from both rejected and accepted papers, The dataset contains peer review report characteristics; reviewer recommendation, reviewer gender, continental region of the reviewer, journal discipline (health and medical sciences-HMS; life sciences-LS; physical sciences-PS; social sciences and economics-SSE) and JIF quartile. The database was filtered to include only review reports from research articles and complete records that contained all of the aforementioned variables, which resulted in a total sample of 496,928 peer-review reports. In the anonymized dataset, we provide (https://osf.io/d34b9/?view_only=), which does not contain texts of peer review, we also describe each statistical glossary word category as a binary variable (if the word from the category was found in the report). Also, we provide data for the two main categories, Methodology and Statistics, again as a binary variable (No/Yes). The methodology is labeled as "Yes" if there was at least one word found from the following categories: "Hypothesis", "Study

design", "Sampling", "Procedures", "Biases", and "Reporting guidelines". Statistics is labelled as "Yes" if there is at least one-word present from any of the following categories: "Distribution", "Meta-analysis", "Variable", "Accuracy and precision", "Parametric", "Non-parametric", "Two group comparison", "Multi-group", "Diagnostics", "Association", "Effects", "Data presentation", and "Transparency". Finally, the word count of peer review was standardized on a scale from 0 to 100. All of the reports were from the first round of reviews.

3.4. Gender of the reviewers

The PEERE dataset does not contain a gender variable, so we implemented a procedure to estimate gender based on the name. Our procedure for gender guessing followed a two-step disambiguation algorithm validated in previous research (Bravo et al., 2019; Buljan et al., 2020; Squazzoni, 2021) that has been demonstrated to be >95% accurate when classifying the names of academics from multiple origins. First, we queried the Python package gender-guesser about the first names and countries of origin. For names classified by gender-guesser as 'mostly_male', 'mostly_female', 'andy' (androgynous), or 'unknown' (name not found), we queried GenderAPI (https://gender-api.com/). This procedure allowed us to guess the gender of 95.4% of academics in our sample, the remaining 4.6% of academics were assigned an unknown gender (this proportion refers to the original dataset, we give access to the dataset used in the study). Note that this level of gender guessing is consistent with the non-classification rate for names of academics in previous research (Santamaria & Mihaljević, 2018). This is a standard disambiguation algorithm that can also deal with the complexity of Asian names and that ensures a rate of misclassified names lower than 5%, by applying the optimal values found in benchmark 2 (Santamaría and Mihaljević, 2018).

3.5. Length of the reports and deviations from the protocol

In this study, we decided to include all available reviews of research articles from the PEERE database. The inclusion of all reviews deviates from the pre-registration, where we arbitrarily decided that reviews with less than 50 words were likely, not actual reviews. However, based on our later observations, we saw that the shorter reports also contain issues identified by peer reviews. The proportion of peer review reports with less than 50 words was substantial, and the elimination of those reports would present a significant dropout in the data. However, we also compared the results both with and without the <50-word peer reviews and the results were largely similar. In the final sample, the median length of the reports was 196 words (interquartile range 108 to 330, min-max 16-720). In the analysis, due to the skewness of the distribution, we applied a min-max normalization using the following formula: $(S–m)/(M–m)\times100$, where S is the number of words in the longest report, m is the number of

words in the shortest review in the dataset, while *M* is each specific number of words per review (Donders et al., 2006). After this normalization, we standardized the length of the reports on a scale from 0 to 100. The median standardized length was 26 (interquartile range (IQR)=13-45), and we used this normalized variable in the regression models. In the pre-registration, one of the intended predictors was the type of peer review (open vs. blind peer review). However, during the analysis of the data, it was observed that that would lead to a great decrease in the number of reviews since we had information about the type of peer review for about half of peer review reports.

3.6. Validation of the glossary

To validate that the glossary is assessing the intended concepts, we extracted anonymized sentences from the peer review reports with low or high aggregated scores. The sentences were extracted separately for the aggregated methodology and statistics variables (see Appendix) across different review characteristics. We extracted sentences across different disciplines, reviewer recommendations, reviewer gender, impact factor quartile, and continental region which would have extremely high methodology/statistics scores and those were qualitatively assessed by the authors for the validity of the glossary. We provided examples in the Appendix as evidence that the texts with high methodology/statistics scores contain text that discusses the mentioned issues, compared to (also provided) reviews where methodological/statistical terms were not contained. We believed that the differences between texts with higher and lower scores would be more recognizable than in cases where the scores are lower and the differences are more nuanced.

3.7. Comparison of term frequency and prediction of term occurrence

We analyzed the data on reviewer characteristics (gender, continental region), article recommendation, and journal (journal area, JIF quartile: first, second, third quartile, and Non-Indexed (NI) journals) and presented them as frequencies and percentages. We assessed the prevalence of methodological and statistical issues addressed separately. Mixed effects logistic regression was used to analyze which reviewer and journal characteristics predict the occurrence of methodology or statistical terms, and the results are presented as odds ratios with 95% confidence intervals (CI), while the journal was entered as the random factor. We opted for a binary approach instead of using the frequencies as continuous variables because the distribution of terms was highly skewed and we would not be able to satisfy testing assumptions. Therefore, a binary Aggregated Methodology variable was created (0 – no terms related to the methodology were mentioned in the peer review report, and 1 – at least one word related to methodology from the glossary was mentioned in the peer review), as well as a binary Aggregated Statistics variable (0 – no terms related to the statistics were mentioned in the peer review

report, and 1 – at least one word related to statistics from the glossary was mentioned in the peer review). Furthermore, to understand what predicts different reviewer recommendations, ordinal logistic regression was used to assess whether the developed categories predict the different reviewer recommendations (Reject = 1, Major revision = 2, Minor revision = 3, and Accept = 4). The analysis was done using the R statistical program (https://www.r-project.org/) and the JAMOVI package for statistical analysis (https://www.jamovi.org/).

4. Results

4.1. Validation of the glossary

As the first aim of this study, we constructed and validated a glossary of terms related to methodological and statistical aspects used in reviews by using a dictionary-based approach. The final glossary contained 1036 terms and can be found on the project website on OSF (https://osf.io/d34b9/?view_only). 54 reviews with the high methodology or statistical aggregated scores were selected for validation. A sample of anonymized reviews can be found in the Supplement.

The reviews with high methodology scores were focused on issues related to research performance and planning, while reviews with high statistical scores focused on issues related to data analysis and graphical presentation.

This was in line with the results from the complete dataset, as the most prevalent methodology categories were “Study design” and “Hypothesis” (Figure 1). The most frequent statistical categories were “Data presentation” and “Parametric statistics” (Figure 1).

[Figure 1 insert here]

4.2. Characteristics of review reports in the PEERE dataset

The majority of the reviewers were male and resided in either Europe or North America (Table 1). The journals were predominantly from the first JIF quartile (Table 1) and PS. Among reviewer recommendations, “accept” was by far the least frequent recommendation (Table 1). The smallest discrepancy in the proportion of male and female reviewers was in SSE, while in other areas the proportion of male reviewers exceeded 75% (Table 1). PS had, compared to other research disciplines, the greatest proportion of reviewers coming from Asian countries, and the lowest proportion coming from North America (Table 1). PS also had the lowest proportion of acceptance recommendations compared to

the other scientific disciplines, while the lowest proportion of rejected recommendations was from SSE (Table 1). Finally, in PS, the vast majority of the journals were from the first JIF quartile, and none were from the third JIF quartile (Table 1). For other disciplines, there was more dispersion across JIF quartiles, although in all research disciplines, most peer review reports were for journals in the first JIF quartile and the least for the journals in the third quartile (Table 1).

[Insert Table 1 here]

4.3. Methodological terms

Terms related to the methodological aspects of reported studies were found in 138,404 (26.8%) peer review reports. A review was more likely to use words related to methodological aspects when the reviewer lived in North America or Oceania, and when the recommendation was to reject the article or recommend major revisions. The reports with methodological words were more frequent in HMS and SSE compared to the LS or PS, but JIF quartile was not predictive for the occurrence of methodological words (Table 2). An additional descriptive analysis showed that journals from the social sciences had the highest proportion of review reports with words from the subcategories "Procedures" and "Study design" compared to other fields (Appendix, Figure A).

[Insert Table 2 here]

4.4. Statistical terms

Terms related to statistics were found in 351,349 (70.7%) of the reviews. The occurrence of statistics-related terms was more likely when the reviewer was female, for reviews recommending major revisions, when the journal was from HMS, and when the review text was longer (Table 3). There were also geographical differences, with the lowest occurrence associated with reviewers living in Africa or Asia, compared to other global regions (Table 3). Although general scores indicated that statistical terms were most prevalent in HMS and PS, the analysis of subcategories showed that the distribution of subcategories varied across research disciplines (Appendix, Figure B). When observing the proportions of the words from different statistical subcategories, HMS did not dominate in any of the compared subcategories. Words from the subcategories "Association", "Meta-analysis", "Parametric descriptive statistics", and "Variable characteristics" were most prevalent in reviews from SSE. "Two-group comparison" and "Multi-group comparison" were most prevalent in reviews from LS, whereas "Data presentation" was most prevalent in reviews from PS (Appendix, Figure B).

[Insert Table 3 here]

4.5. Prediction of reviewers' recommendations

To dig deeper into what predicts different reviewer recommendations, we performed an ordinal logistic regression with review recommendation as the dependent variable and different subcategories from both Statistics and Methods as the dependent variables. Besides reviewer and journal characteristics, a large number of methodology and statistics categories were predictive (Deviance residuals Md=-0.06,

IQR -0.93 to 0.77; Akaike Information Criterion=1316428; Bayesian Information Criterion=1316806; Table 4), but the proportion of explained variance was low (McFadden R^2=0.015). Review reports that contained words related to study design, hypothesis, sampling, two-group comparison, and transparency were related to less favorable review recommendations. Meanwhile, reviews that contained words related to reporting guidelines, nonparametric statistics, effect sizes, and data presentation were related with more favorable review recommendations (Table 4).

[Insert Table 4 here]

5. Discussion

Our study showed that an analysis of peer review reports based on a glossary of statistical and methodological terms may be useful for studying the content of peer review reports and for understanding how authors can be helped to write manuscripts in a way to increase the clarity of their methodological and analytic approach. To develop a method that could assess methodological and statistical issues in peer reviews, in the first part of the study, we constructed a glossary of terms that can be used in content assessment when examining the prevalence of methodological and statistical issues in peer review reports. We applied the glossary to a large dataset of peer review reports and found that a mention of methodological issues was related to more "reject" recommendations, and longer review reports, and was more common when the reviewer was from North America or Oceania and in journals from the SSE or HMS. On the other hand, a mention of statistical issues was more common if the reviewer was female, when the recommendation was a major revision, when the journal was from HMS or PS, and when reviews were longer. However, when entered into the model that predicted a reviewer's recommendation, glossary subcategories explained a very small proportion of variance, indicating that other factors contribute to the final reviewer's recommendation.

The results of this study should be interpreted in light of several limitations. First, when adopting a glossary approach in quantitative textual analysis, the results only indicate the number of specific words used in the reviews. It does not provide the context of the mention, so it is not clear whether the words identified in the reports represent positive (i.e., commending the use of a technique) or negative (i.e., critiquing the use of a technique) feedback. Therefore, it is reasonable to question whether the frequency of these mentions reflects underlying improvements that need to be made to the research. However, due to the nature of the feedback provided in peer review reports and the anonymized reviews we present in the Appendix, we consider that the identification of words that occur in reviews will most often be associated with potential issues in the article. The prevalence of identified words differed between research disciplines, indicating that reviewers from different disciplines focus on different areas when it comes to methodological and statistical issues. In the logistic regression, we found that methodological words in our glossary were most prevalent in the SSE disciplines, while statistical words were more prevalent in HMS. Another potential limitation is that we cannot distinguish whether the reviewers from SSE identified more methodological issues, or whether our glossary was more suited to detecting the specific terms used in the SSE. When constructing the glossary, we tried to implement broad terms, related to different research fields (Appendix, Table S4). We also found a relatively high absolute number of terms in all research disciplines explored, suggesting that the glossary is suitable for different research disciplines. Finally, it needs to be acknowledged that our glossary may be suitable for quantitative

methodology only, due to the inclusion of words and terms specifically related to that methodological approach, so our glossary may not applicable to qualitative studies and reviews. Also, the terms were categorized initially by the reporting guidelines categories, and refined further by reviewers' experience. It is expected that in the future glossary will be further expanded and updated according to the specific user needs, in order to capture more terms from various subfields.

Despite the limitations, there are several theoretical and practical implications of the results from our study. The first theoretical implication is that we showed that it is possible to create a comprehensive glossary for statistical and methodological terms. In general, we found that more review reports mentioned statistical issues compared to methodological issues. The possible conclusion is that reviewers are either more focused on statistical analysis in the reviews and require additional analyses and revisions, or that authors are simply more fallible when it comes to statistical analysis. As we developed more glossary items related to statistics, it is also possible that our glossary captures a greater range of statistical concepts compared to methodological concepts. Nonetheless, methodological words were related to more "reject" recommendations, while statistical words were more related to revision recommendations. This is intuitive as many statistical issues could be corrected during revisions, while methodological issues may require a full revision of the study plan/protocol, as has been shown by our previous study on a small sample (Garcia-Costa et al., 2022). Also, our model had lower predictive performance compared to a model from a previous study which attempted to predict retractions based on peer reviews (Zheng et al., 2023), but the sample size in that study was small (N=260) and compared only binary outcomes and therefore not suitable for comparison. Consequently, the practical implication of the results from our study is that studies should be reviewed at the protocol stage, to reduce research waste and errors, and increase the probability of acceptance after research has been performed. Examples of such results-free peer review exist in social sciences (Findley et al., 2016; Chalmers & Tzavella, 2021), as well as in evidence synthesis in medical sciences, such as protocols for Cochrane systematic reviews of health interventions (e.g. Higgins et al., 2023).

Another theoretical implication is related to the prevalence of the glossary terms. The most frequent categories mentioned under the statistical glossary were related to data presentation, parametric statistics, and variable characteristics. Although it could be expected that the majority of scientific articles would have those statistical concepts included in their analysis, and more specific concepts (e.g. meta-analysis) would be significantly less prevalent in research articles, there is also a potential alternative explanation for this finding. A previous study found that only about one-fifth of biomedical journals use specialized statistical review (Hardwicke & Goodman, 2020). There is not much evidence for this in other disciplines, and it is possible that this finding could also be different across disciplines since social

sciences could be primarily researchers and therefore better versed in statistics, in contrast to medical researchers who may be primarily health, while a lot of biomedical authors are primarily practitioners and less focused on research statistics. However, because the majority of reviewers are not statistical experts, they may focus on only basic aspects related to the data in the reviewed articles. Our study showed significant variation across disciplines in the proportion of review reports with words from different subcategories. The most substantial differences were found in the subcategories “Procedures” and “Study design”, with SSE having the highest prevalence of reviews containing words from those categories. A potential explanation is that the reviewers identify more methodological issues in social sciences, or the peer reviewers in social sciences are more focused on the subcategories in question compared to other disciplines. The same applies to statistics. There is a possibility that some aspects of the statistical analysis are more relevant to some fields than in other fields (e.g., data presentation in physics, compared to other disciplines), and therefore the reviewers are more focused on them in their peer review reports. From the current results, we can conclude that specific categories were present in different disciplines, like data presentation, parametric statistics, and variable characteristics. However, the prevalence of other categories differed across disciplines, except for HMS where all categories were equally prevalent. It must also be noted that we began with categories related to medical research in the development of the glossary, and there is a possibility that the current version of the glossary is more suitable for HMS. An in-depth analysis of the peer review reports and the actual content of the articles would be needed to address this possibility.

Another practical implication relates to the glossary itself, which can be modified (expanded or shortened) according to the user's needs, in case users want to adopt the dictionary-based approach in the assessment of a large set of research articles or peer reviews. As previously noted, this method could be a complementary approach to deeper methods of analysis, which are frequently difficult to achieve in large datasets. Furthermore, it is to be seen how AI tools can be utilized in combination with the glossary since those tools progress at tremendous speed, and it remains to be seen how those tools will be integrated into the review process (e.g. Checco et al., 2021). Another practical implication was that a higher JIF quartile was not associated with a higher probability of finding methodological or statistical terms in peer-review reports, which means that all categories of JIF had the same probability of methodological and/or statistical issues. Unfortunately, we cannot know whether studies submitted to journals with different JIFs generally differ in quality or whether journals with high JIFs can recruit more expert reviewers. Therefore, it is expected that those reviewers would provide more substantial feedback related to the methods and analyses in the study, and potentially hold the articles to a higher standard. This cannot be tested in the current dataset as we have no information about the expertise of the reviewers. However,

investigating this association may be a fruitful path for further study. The recommendation is, if possible, also to investigate the review content based on the other characteristics which predict journal impact, like journal h-index (Mingers, Macri & Petrovici, 2012). Finally, since the specific glossary dimensions explained a low proportion of variance of reviewer recommendations, future studies should focus on the research of which are factors that have more significant effect on reviewer recommendations.

6. Conclusions

Our study is currently the first to construct a comprehensive glossary of methodological and statistical terms and apply it to a large sample of peer review reports. Some of these results are in line with studies conducted on published articles, which did not include rejected submissions (Bolt et al., 2021). We find that methodological issues mentioned in review reports are more likely to result in a "reject" decision, while reviews mentioning statistical analysis can be corrected as long as methods are well planned. Due to the large discrepancy in the prevalence of different categories, a machine-learning approach would likely not be suitable for the identification of all categories/issues.

Further studies might apply the glossary in different contexts and fields to increase external validity in the identification of methodological and statistical issues. There are already algorithms that use text mining to determining the risk of unsuccessful completion of a scientific study, for which our dictionary could be complementary (Follet, Gelleta & Laugerman, 2019). Potential applications also include the analysis of online discussions, informal communications, and comments in scientific communities, which would present complementary data to the current data on peer review reports.

The newly constructed glossary can be used in the detection of statistical and methodological words in research-related reports (peer review reports, research articles, conference proceedings). It is also openly available (https://osf.io/d34b9/files/osfstorage), and suitable for future expansion and improvement. Furthermore, as shown here, the most prevalent statistical and methodological categories are equally distributed across disciplines, indicating that reviewers are either more focused on descriptive statistics, or their statistical knowledge is limited to basic statistics. Finally, while methodological issues were predicted by rejection recommendation, statistical issues were better predicted by revisions, indicating that methodological issues were potentially the reason for rejection and statistical issues for revisions, but this needs to be confirmed by an in-depth analysis.

Abbreviations:

JIF- Journal Impact factor

CI- Confidence interval

HMS- Health and medical sciences

LS- Life sciences

PS- Physical sciences

SSE- Social sciences and economics

NI- Not indexed

Declaration of interests:

Authors have no competing interests to declare.

Declaration of use of generative AI:

The authors state that generative AI was not used in the preparation of this manuscript.

Acknowledgments: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract; and decision to submit the abstract for presentation.

Funding

This study was funded by the Croatian Science Foundation "Professionalism in Health–Decision making in practice and research" under grant agreement IP-2019-04-4882. Richard A. Klein received funding from a Replication Studies grant (401.18.053) from the Netherlands Organization for Scientific Research, and a Consolidator Grant (IMPROVE) from the European Research Council (grant 726361). Francisco Grimaldo and Daniel Garcia-Costa are partially supported by the Spanish Ministry of Science, Innovation and Universities, the Spanish State Research Agency, and the European Regional Development Fund under project RTI2018-095820-B-I00.

Data statement

The journal dataset required a data sharing agreement to be established between authors and publishers. A protocol on data sharing entitled 'TD1306 COST Action New frontiers of peer review (PEERE) PEERE policy on data sharing on peer review' was signed by all partners involved in this research on 1 March 2017, as part of a collaborative project funded by the EU Commission. The protocol established rules and practices for data sharing from a sample of scholarly journals, which included a specific data management policy, including data minimization, retention and storage, privacy impact assessment, anonymization, and dissemination. The protocol required that data access and use were restricted to the authors of this manuscript and data aggregation and report were done in such a way to avoid any identification of publishers, journals or individual records involved. The protocol was written to protect the interests of any stakeholder involved, including publishers, journal editors and academic scholars, who could be potentially acted by data sharing, use and release. The full version of the protocol is available on the peere.org website. To request additional information on the dataset and for any claim or objection, please contact the PEERE data controller at info@peere.org.

References

Andaur Navarro, C. L., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ (Clinical research ed.), 375*, n2281. doi:10.1136/bmj.n2281

Bharti, P., Ghosal, T., Agarwal, M., & Ekbal, A. (2022). *A Method for Automatically Estimating the Informativeness of Peer Reviews.* Paper presented at the Proceedings of the 19th International Conference on Natural Language Processing (ICON).

Bolt, T., Nomi, J. S., Bzdok, D., & Uddin, L. Q. (2021). Educating the future generation of researchers: A cross-disciplinary survey of trends in analysis methods. *Plos Biology, 19*(7). doi: 10.1371/journal.pbio.3001313

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications, 10*(1), 322. doi: 10.1038/s41467-018-08250-2

Buljan, I., Garcia-Costa, D., Grimaldo, F., Squazzoni, F., & Marušić, A. (2020). Large-scale language analysis of peer review reports. *eLife, 9*, e53249. doi: 10.7554/eLife.53249

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*(1), 29-42. doi: 10.1038/s41562-021-01193-7

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications, 8*(1), 25. doi: 10.1057/s41599-020-00703-8

Chubb, J., Cowling, P., & Reed, D. (2021). Speeding up to keep up: exploring the use of AI in the research process. *AI & SOCIETY* , 37(4), 1439-1457. doi: 10.1007/s00146-021-01259-0

Deng, Q., Hine, M. J., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: a design science approach. *Journal of international technology and information management, 27*(3), 119-159. doi:10.58729/1941-6679.1376

Dimity, S. (2022). Peer reviewers equally critique theory, method, and writing, with limited effect on the final content of accepted manuscripts. *Scientometrics, 127*(6), 3413-3435. doi: 10.1007/s11192-022-04357-y

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091. doi: 10.1016/j.jclinepi.2006.01.014

Findley, M. G., Jensen, N. M., Malesky, E. J., & Pepinsky, T. B. (2016). Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study. *Comparative Political Studies, 49*(13), 1667-1703. doi: 10.1177/0010414016655539

Follett, L., Geletta, S., & Laugerman, M. (2019). Quantifying risk associated with clinical trial termination: A text mining approach. *Information Processing & Management, 56*(3), 516-525. doi: 10.1016/j.ipm.2018.11.009

Fox, C. W. (2017). Difficulty of recruiting reviewers predicts review scores and editorial decisions at six journals of ecology and evolution. *Scientometrics, 113*(1), 465-477. doi: 10.1007/s11192-017-2489-5

Garcia-Costa, D., Squazzoni, F., Mehmani, B., & Grimaldo, F. (2022). Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ, 10*, e13539. doi: 10.7717/peerj.13539

Garcia-Costa, D., Forte, A., Lòpez-Iñesta, E., Squazzoni, F., & Grimaldo, F. (2022). Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals. *Royal Society open science*, *9*(9), 210681. doi: 10.1098/rsos.210681

Gender-API. Available at: https://gender-api.com/ Retrieved on 16th December 2022.

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PloS one, 17*(1), e0259238. doi: 10.1371/journal.pone.0259238

H2O.ai Available at: https://www.h2o.ai/ Retrieved on 16th December 2022.

Hardwicke, T., & Goodman, S. (2020). How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey. *PloS one, 15*, e0239598. doi: 10.1371/journal.pone.0239598

Han, R., Zhou, H., Zhong, J., & Zhang, C. (2022). Characterizing Peer Review Comments of Academic Articles in Multiple Rounds. *Proceedings of the Association for Information Science and Technology*, *59*(1), 89-99.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2023). *Cochrane Handbook for systematic reviews of interventions* (Version 6.4, updated August 2023). Cochrane. https://www.training.cochrane.org/handbook

Holosko MJ, Thyer BA. Pocket Glossary for Commonly Used Research Terms. Thousand Oaks, CA: SAGE Publications, Inc.; 2011. doi:10.4135/9781452269917

Horbach, S. P. J. M., Oude Maatman, F. J. W., Halffman, W., & Hepkema, W. M. (2022). Automated citation recommendation tools encourage questionable citations. *Research Evaluation, 31*(3), 321-325. doi: 10.1093/reseval/rvac016

Huber, J., Inoua, S., Kerschbamer, R., König-Kersting, C., Palan, S., & Smith, V. L. (2022). Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences of the United States of America, 119*(41), e2205779119. doi:10.1073/pnas.2205779119

588 Jacobucci, R., Ammerman, B. A., & Wilcox, K. T. (2021). The use of text-based responses to improve
589 our understanding and prediction of suicide risk. *Suicide and Life-Threatening Behavior, 51*(1), 55-
590 64. doi: 10.1111/sltb.12668
591 Kilicoglu, H., Rosemblat, G., Hoang, L., Wadhwa, S., Peng, Z. S., Malicki, M., . . . ter Riet, G. (2021).
592 Toward assessing clinical trial publications for reporting transparency. *Journal of Biomedical*
593 *Informatics, 116*. doi: 10.1016/j.jbi.2021.103717
594 Lauer, M., Constant, S., & Wernimont, A. (2023). Using AI in Peer Review Is a Breach of
595 Confidentiality. National Institutes of Health: Office of Extramural Research. Accessed on
596 September 17th 2023. Available from: https://nexus.od.nih.gov/all/2023/06/23/using-ai-in-peer-
597 review-is-a-breach-of-confidentiality/
598 Meng, J. (2023). Assessing and predicting the quality of peer reviews: a text mining approach. The
599 Electronic Library, 41(2/3), 186-203. doi:10.1108/EL-06-2022-0139
600 Mingers, J., Macri, F., & Petrovici, D. (2012). Using the h-index to measure the quality of journals in the
601 field of business and management. *Information Processing & Management, 48*(2), 234-241. doi:
602 10.1016/j.ipm.2011.03.009
603 Mpouli, S., Beigbeder, M., & Largeron, C. (2020). Lexifield: a system for the automatic building of
604 lexicons by semantic expansion of short word lists. *Knowledge and Information Systems, 62*(8),
605 3181-3201. doi: 10.1007/s10115-020-01451-6
606 Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., . . . Aczel, B. (2022).
607 A community-sourced glossary of open scholarship terms. *Nature Human Behaviour, 6*(3), 312-
608 318. doi: 10.1038/s41562-021-01269-4
609 Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of
610 LIWC2015. 2015. Austin, TX: University of Texas at Austin. doi: 10.15781/T29G6Z
611 Perković Paloš, A., Mijatović, A., Buljan, I., Garcia-Costa, D., Álvarez-García, E., Grimaldo, F., &
612 Marušić, A. (2023). Linguistic and semantic characteristics of articles and peer review reports in
613 Social Sciences and Medical and Health Sciences: analysis of articles published in Open Research
614 Central. *Scientometrics, 128*(8), 4707-4729. doi: 10.1007/s11192-023-04771-w
615 Petchiappan, R., James, K., Plume, A., Tsakonas, E., Marusic, A., Malički, M., . . . Mehmani, B. (2022).
616 Analysing Elsevier Journal Metadata with a New Specialized Workbench inside ICSR Lab. *SSRN*
617 *Electronic Journal*. doi: 10.2139/ssrn.4211833
618 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical
619 Computing, Vienna, Austria. 2021.

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science, 6*(1), 28. doi: 10.1140/epjds/s13688-017-0121-9

Reveilhac, M., & Morselli, D. (2022). Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data. *Political Research Exchange, 4*(1). doi: 10.1080/2474736x.2022.2029217

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science, 4*, e156. doi: 10.7717/peerj-cs.156

Shopovski, J., Bolek, C., & Bolek, M. (2020). Characteristics of Peer Review Reports: Editor-Suggested Versus Author-Suggested Reviewers. *Science and Engineering Ethics, 26*(2), 709-726. doi: 10.1007/s11948-019-00118-y

Simera, I., Moher, D., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). A catalogue of reporting guide-lines for health research. *European Journal of Clinical Investigation, 40*(1), 35-53. doi:10.1111/j.1365-2362.2009.02234.x

Sizo, A., Lino, A., Reis, L. P., & Rocha, Á. (2019). An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management, 46*, 286-293. doi: https://doi.org/10.1016/j.ijinfomgt.2018.07.002

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., Bravo, G., Cowley, S., Dignum, V., Dondio, P., Grimaldo, F., Haire, L., Hoyt, J., Hurst, P., Lammey, R., MacCallum, C., Marušić, A., Mehmani, B., Murray, H., Nicholas, D., … Willis, M. (2020). Unlock ways to share data on peer review. *Nature, 578*(7796), 512–514. doi:10.1038/d41586-020-00500-y

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., . . . Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances, 7*(2), eabd0299. doi:10.1126/sciadv.abd0299

Stark PB. Glossary of Statistical Terms. Available at: https://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm. Last accessed: 23rd March 2021.

Sun, Z., Clark Cao, C., Ma, C., & Li, Y. (2023). The academic status of reviewers predicts their language use. *Journal of Informetrics, 17*(4), 101449. doi:10.1016/j.joi.2023.101449

The jamovi project (2022). jamovi (Version 2.3). Retrieved from https://www.jamovi.org

van Atteveldt, W., van der Velden, M., & Boukes, M. (2021). The Validity of Sentiment Analysis:Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures, 15*(2), 121-140. doi: 10.1080/19312458.2020.1869198

Thelwall, M. (2023). Journal and disciplinary variations in academic open peer review anonymity, outcomes, and length. *Journal of Librarianship and Information Science, 55*(2), 299-312.

Urban, L., De Niz, M., Fernández-Chiappe, F., Ebrahimi, H., Han, L. K. M., Mehta, D., . . . Yahia Mohamed Elkheir, L. (2022). eLife's new model and its impact on science communication. *eLife, 11*, e84816. doi: 10.7554/eLife.84816

Weiss, M., Nair, L. B., Hoorani, B. H., Gibbert, M., & Hoegl, M. (2023). Transparency of reporting practices in quantitative field studies: The transparency sweet spot for article citations. *Journal of Informetrics, 17*(2), 101396. doi:10.1016/j.joi.2023.101396

Zheng, X., Chen, J., Tollas, A., & Ni, C. (2023). The effectiveness of peer review in identifying issues leading to retractions. *Journal of Informetrics, 17*(3), 101423. doi:10.1016/j.joi.2023.101423

Table 1

Characteristics of peer review reports analysed in the study

	Research discipline (n, %)			
Variable	HMS (n=139501)	LS (n=69373)	PS (n=269333)	SSE (n=18721)
Gender:				
Female	31010 (22.2)	16489 (23.8)	55125 (20.5)	7795 (41.6)
Male	108491 (77.8)	52884 (76.2)	214208 (79.5)	10926 (58.4)
Continental region:				
Africa	1656 (1.2)	681 (1.0)	4130 (1.5)	218 (1.2)
Asia	18263 (13.1)	5601 (8.1)	69709 (25.9)	1277 (6.8)
Europe	46489 (33.3)	30796 (44.4)	119787 (44.5)	6771 (36.2)
North America	63803 (45.7)	25666 (37.0)	56440 (21.0)	9022 (48.2)
Oceania	6528 (4.7)	4441 (6.4)	8983 (3.3)	1283 (6.9)
South America	2762 (2.0)	2188 (3.2)	10284 (3.8)	150 (0.8)
Recommendation:				
Accept	19889 (14.3)	5269 (7.6)	12270 (4.6)	1908 (10.2)
Major revision	46078 (33.0)	22396 (32.3)	102226 (38.0)	7361 (39.3)
Minor revision	30581 (21.9)	23880 (34.4)	74152 (27.5)	2489 (13.3)
Reject	42953 (30.8)	17828 (25.7)	80685 (30.0)	6963 (37.2)
JIF quartile:				
NI	44514 (31.9)	6952 (10.0)	4951 (1.8)	3985 (21.4)
Q1	44761 (32.1)	32855 (47.4)	226948 (84.3)	10939 (58.4)
Q2	27430 (19.7)	22671 (32.7)	37434 (13.9)	3099 (16.6)
Q3	22796 (16.3)	6895 (9.9)	0 (0.0)	698 (3.7)

Abbreviations: HMS-Health and Medical Sciences, LS-Life Sciences, PS-Physical Sciences, SSE-Social Sciences and Economics, JIF-Journal impact factor, NI-not indexed

Table 2

Mixed effects logistic regression model for prediction of methodological words in the review while controlling for journal

Predictor	Odds ratio	95% CI	P
Intercept	0.32	0.21 to 0.50	<0.001
Recommendation			
Ref ("accept")			
Minor revision	1.24	1.20 to 1.28	<0.001
Major revision	1.56	1.51 to 1.61	<0.001
Reject	1.68	1.62 to 1.74	<0.001
Continental region:			
Ref (Africa)			
Asia	0.91	0.85 to 0.97	0.007
Europe	1.00	0.93 to 1.07	0.924
North America	1.17	1.09 to 1.25	<0.001
Oceania	1.11	1.03 to 1.20	0.004
South America	0.98	0.91 to 1.20	0.689
Gender:			
Ref (female)			
Male	0.97	0.96 to 0.99	0.004
Research discipline:			
Ref (HMS)			
LS	0.23	0.15 to 0.36	<0.001
PS	0.34	0.23 to 0.52	0.003
SSE	1.11	0.44 to 2.74	0.919
JIF quartile:			
Ref (NI)			
Q1	1.01	0.57 to 1.82	0.978
Q2	0.90	0.50 to 1.63	0.874
Q3	0.77	0.33 to 1.81	0.817
Word count (transformed on the scale from 0-100)	1.03	1.03 to 1.03	<0.001

Abbreviations: JIF-Journal impact factor, NI-not indexed, HMS-Health and Medical Sciences, LS-Life Sciences, PS-Physical Sciences, SSE-Social Sciences and Economics, IF-impact factor, NI-not indexed

Table 3

Mixed effects logistic regression model for prediction of statistical words in the review while controlling for journal

Predictor	Odds ratio	95% CI	P
Intercept	3.04	2.70 to 3.44	<0.001
Recommendation			
Ref ("accept")			
Minor revision	1.61	1.56 to 1.65	<0.001
Major revision	1.72	1.66 to 1.77	<0.001
Reject	1.29	1.24 to 1.32	<0.001
Continental region:			
Ref (Africa)			
Asia	1.03	0.98 to 1.10	0.222
Europe	1.11	1.05 to 1.17	<0.001
North America	1.19	1.12 to 1.26	<0.001
Oceania	1.16	1.08 to 1.27	<0.001
South America	1.26	1.17 to 1.35	<0.001
Gender:			
Ref (female)			
Male	0.90	0.88 to 0.92	<0.001
Research discipline:			
Ref (HMS)			
LS	0.64	0.53 to 0.78	<0.001
PS	0.87	0.72 to 1.05	0.145
SSE	0.50	0.34 to 0.74	<0.001
J IF quartile:			
Ref (NI)			
Q1	1.15	0.90 to 1.46	0.269
Q2	0.98	0.76 to 1.27	0.915
Q3	1.09	0.74 to 1.58	0.664
Word count (transformed on the scale from 0-100)	1.06	1.06 to 1.07	<0.001

Abbreviations: JIF-Journal impact factor, NI-not indexed, HMS-Health and Medical Sciences, LS-Life Sciences, PS-Physical Sciences, SSE-Social Sciences and Economics, IF-impact factor, NI-not indexed

Table 4

Ordinal regression on the entire sample, with the reviewer's recommendation as a predictor

Levels	Odds ratio	95% CI	P[a]
Intercept	10.94	10.68 to 11.20	<0.001
Gender:			
Ref (female)			
Male	0.94	0.93 to 0.95	<0.001
Research discipline:			
Ref (Health and Medical Sciences)			
Life sciences	1.06	1.05 to 1.07	<0.001
Physical sciences	0.81	0.81 to 0.82	<0.001
Social sciences and Economics	0.85	0.84 to 0.90	<0.001
Journal IF quartile:			
Ref (NI)			
Q1	1.21	1.19 to 1.22	<0.001
Q2	1.07	1.06 to 1.08	<0.001
Q3	0.90	0.88 to 0.91	<0.001
Continental region:			
Ref (Africa)			
Asia	0.82	0.81 to 0.84	<0.001
Europe	0.95	0.93 to 0.97	<0.001
North America	1.03	1.00 to 1.05	0.009
Oceania	0.98	0.95 to 1.00	0.125
South America	0.96	0.94 to 0.99	0.003
Study design	0.95	0.94 to 0.95	<0.001
Hypothesis	0.95	0.94 to 0.95	<0.001
Sampling	0.95	0.94 to 0.96	<0.001
Procedures	0.98	0.97 to 1.00	0.012
Biases	0.99	0.98 to 1.02	0.930
Reporting guidelines	1.09	1.05 to 1.14	<0.001
Distribution	1.01	0.98 to 1.01	0.534
Meta-analysis	0.99	0.98 to 1.01	0.269
Variables	0.98	0.98 to 0.99	0.010
Accuracy	1.04	1.01 to 1.06	0.008
Parametric	0.96	0.95 to 0.97	<0.001
Nonparametric	1.08	1.07 to 1.10	<0.001
Two group comparison	0.93	0.92 to 0.94	<0.001
Multiple group comparison	0.99	0.97 to 0.99	0.003
Diagnostics	0.99	0.98 to 1.00	0.042
Association	1.00	0.99 to 1.02	0.767
Effects	1.05	1.03 to 1.06	<0.001
Data presentation	1.05	1.05 to 1.06	<0.001
Transparency	0.94	0.93 to 0.94	<0.001
Word count	0.99	0.99 to 0.99	<0.001

Abbreviations: JIF-Journal impact factor.

[a]The order of the reviewer recommendations was: Reject-1, Major revision-2, Minor revision-3, Accept-4.