Scientometrics (2026) 131:59–92
https://doi.org/10.1007/s11192-026-05539-8

Automatically investigating scientific discussions in peer review reports based on conformity score metrics

Qiao Sun[1] · Dahai Liu[1] · Haochen Dong[1] · Shuangfei Li[1]

Received: 11 January 2025 / Accepted: 5 January 2026 / Published online: 22 January 2026
© The Author(s) 2026

Abstract
This study proposes a conformity score metric to facilitate the editorial evaluation of peer review report quality. This metric assesses review reports' conformity with quality standards by identifying and quantifying the contribution of individual sentences within the content. When the conformity score metric is applied to a review report, sentences relating to the methodological aspects of the research reviewed are assigned the highest score relative to sentences relating to the reviewers' discussions of other aspects of the reviewed research. To identify review sentences focusing on different aspects of the reviewed research, this study uses the Long Short-Term Memory (LSTM) model, which classifies sentences from peer review report texts into a number of categories. Based on review report textual data, this study examines the relationship between features that are commonly regarded as indicators of review report quality and conformity scores, which are determined to have a strong correlation with these textual features. Overall, this study proposes a review report quality assessment metric that considers the semantic meaning of review content and emphasizes discussions about the scientific aspects of the reviewed research. The conformity score metric and the LSTM model can thus be integrated into review report examination practices.

Keywords Peer review quality · Text mining · Machine learning

Introduction

As a core aspect of academic publishing, peer review plays a crucial role in controlling research quality and promoting scientific exchange. During the editorial peer review process, the appraisal of scientific contributions is universally accepted as a benchmark for evaluating research quality (Jefferson et al., 2002). From a practical point of view, peer review relies on reviewers scrutinizing manuscripts and writing detailed reports expressing their opinions about the manuscript's quality. The reviewers' comments are expected to promote the overall quality of the research, and their recommendations (e.g., acceptance

✉ Qiao Sun
stevensunss@163.com

[1] MDPI(Beijing) Science & Technology Service Ltd. Co., Poly Metropolitan Floor 9-11, Building 2, Courtyard 4, Guanyinan North Street, Beijing 101101, China

for publication or rejection) should inform editors' decisions when determining whether a submission is suitable for publication in its current iteration. However, the current peer review paradigm features some notable challenges with regard to subjectivity, unpredictable review times, and the variable quality of individual reviews (Tennant & Ross-Hellauer, 2020). These problems threaten the editorial process and they cannot always be easily overcome via training or other similar interventions (Bruce et al., 2016; Enserink, 2001; Houry et al., 2012; Jefferson et al., 2006). These dilemmas are further exacerbated by the increasing number of submissions that a peer review organizer is expected to handle, resulting in mounting pressure on peer review organizers to efficiently screen for high-quality review reports (Sculley et al., 2018). Therefore, in the face of these challenges, determining how to assess review report quality and improving screening efficiency are of the central concern for both academics and practitioners.

High-quality peer review reports should provide constructive advice for authors, allowing them to develop their work. They should also serve as reliable references for editors, assisting in their decision-making when determining whether a manuscript is of a publishable standard (Shattell et al., 2010; Van Rooyen, 2001). To overhaul the peer review system and increase its value, some researchers and practitioners propose certain standards that, when met, ensure reports are of high quality (Falkenberg & Soranno, 2018; Rajesh et al., 2013; Sculley et al., 2018; Superchi et al., 2020; Van Rooyen, 2001). For example, Van Rooyen (2001) proposed a Review Quality Instrument (RQI) that consists of a number of requirements for high-quality review reports (e.g., whether the reviewer discusses the importance of the research questions or the rigorousness of methods employed in the reviewed research).

To the best of our knowledge, these assessment standards evaluate reviewers' discussions of different aspects of the research reviewed equally. However, we argue that a large number of reviewers' discussions about certain aspects (e.g., methodology) indicate the intellectual effort invested, signaling a high-quality review. Superchi et al. (2019) found that many studies describing tools for evaluating review report quality do not address the scientific aspects of such reports, and few evaluation metrics focus on discussions of the methods employed in the articles reviewed. Hopewell et al. (2014) investigated a number of medical journals and found that their review reports contained little discussion of the methods, and many of them had no positive impact on the final publication. Likewise, Henly and Dougherty (2009) found that high-quality reviews generally address the technical aspects of manuscripts. Furthermore, Henly et al. (2010) showed that reviews that emphasize statistical aspects are more useful for editors than those that contain little discussion of statistical procedures.

In a review report, discussions about technical details, such as methodological robustness, are highly beneficial for authors, allowing them to further develop their research. Additionally, the inclusion of a substantial number of technical discussions also reflects a reviewer's commitment to providing a high-quality report. Therefore, in this study, we extend these review conformity standards by highlighting the importance of discussions of technical aspects of the research reviewed. We propose a conformity score metric that evaluates reviewers' comments on a number of aspects of the reviewed research and prioritizes their discussion of technical aspects (e.g., methodologies) in an aggregate scoring

measure,[1] referring to conformity standards proposed in previous studies (Falkenberg & Soranno, 2018; Rajesh et al., 2013; Sculley et al., 2018; Superchi et al., 2020; Van Rooyen, 2001). In addition to identifying reviewers' discissions of methodological aspects, we also examined their discussions about the manuscript's conclusions, contributions, introductions, presentation, references, and results. The finalized conformity score metric is an aggregation of the scores assigned to discussions of individual aspects of the article, among which the methodological aspect receives the largest multiplier.

Callaham et al. (1998) found that editors' subjective ratings of review reports are modestly reliable and mainly relate to their identification of flaws. Therefore, human editors' judgement regarding review quality is, in a sense, reasonable and reliable. However, review assessments can be time-consuming work, and it is hard to maintain consistent standards based due to subjectivity. In recent years, with the rise of Artificial Intelligence (AI), several natural language processing (NLP) technologies have emerged and been integrated into peer review report evaluations (Kuznetsov et al., 2024). These approaches typically use text classifiers to determine whether language content falls within a category indicating quality. Their effective utilization also demonstrates that well-developed AI technologies have justified the feasibility of integrating an automatic peer review report-assessing tool into editorial practices.

In this study, we collect review reports of MDPI publications and identify whether sentences[2] belong to a specific class using text-mining techniques. We manually label sentences in the collected review reports according to seven text classifications (Conclusion, Contributions, Introduction, Methods, Presentation, References, and Results).[3] We train a text classification model using LSTM, which is then used to classify sentences from MDPI review reports. Then, based on this classification, we sum the score for each category[4] for each review report. The total score is referred to as the conformity score and reflects the extent to which a report conforms to established reviewing standards.

In addition to the use of deep learning approaches to investigate review report semantics, some studies use algorithms with rules to extract textual features (Arous et al., 2021; Ausloos et al., 2016; Maddi & Miotti, 2024; Meng, 2023). In a recent study (Meng, 2023), the authors proposed a series of quantitative review report features, including linguistic (e.g., review and sentence length), sentiment (e.g., subjectivity, sentiment polarity) and interaction (e.g., review–abstract similarity) features. Using the above features as model inputs, machine learning models were used to predict review report conformity. The study shows that the features mentioned above are correlated with conformity with publishers'

[1] Evaluating review report quality based on an overall grading score of a review report is a common approach adopted in previous studies (Sizo et al., 2019; Thompson et al., 2016), and can provide review report investigator a general understanding of a review report's quality.

[2] We investigate a review report at the sentence level since a deep understanding about a review report's quality should be facilitated by interpretation of semantic meanings of individual sentences in a review report.

[3] Those categories are consistent with the MDPI review quality examination guidelines and the other similar conformity standards such as RQI.

[4] The score of an individual category is the frequency of sentences in that category and a multiplier, which is determined by the category's importance. For example, the Methods category has the highest multiplier in comparison with others.

standards. To measure the conformity of review reports, researchers asked human participants to grade the reports based on eight criteria proposed by Arous et al. (2021).[5]

As mentioned above, the conformity of a review report's with established standards is related to certain textual features. As such, these features should have predictive power for review report conformity and, if a measure of review report conformity is reasonable, that measure should be correlated with the textual features. To justify the effectiveness of the proposed conformity score metric, we employ a machine learning model that predicts conformity scores using textual features and demonstrate that the predictions are highly correlated with actual scores. Therefore, we argue that our proposed conformity score metric proxies human understandings of review report conformity to a significant extent.

The novelty and highlights of this study are as follows: i) The conformity score metric highlights the importance of technical comments, which are regarded as the most constructive aspects of peer review discussions. To the best of our knowledge, previous studies examining measures of review report quality have not directly addressed discussions about the scientific rigorousness of the reviewed research in detail. ii) The rapid development of large language models also sparks discussions about the feasibility of conducting research reviews based on generative AI tools (D'Arcy et al., 2024; Drori & Te'eni, 2024; Liang et al., 2024; Yuan et al., 2022). Ethical considerations of this application of AI aside, we argue that the conformity score metric's emphasis on technical discussions, which demand high intellectual effort and knowledge from human reviewers, challenges the attempted use of AI tools to produce high-quality reviews. This is because these generative tools (e.g., tools that summarize text) cannot perform well under the evaluative framework of the conformity score metric if they do not address the scientific aspects of the reviewed research. This implication of the evaluative framework is particularly important given the concern that AI generated review reports cannot produce in-depth discussions (Shcherbiak et al., 2024; Zhu et al., 2025). iii) By analyzing the relationship between some textual features and conformity scores, we shed light on discussions about the features of quality review reports. For example, we find that review reports with high conformity scores are generally neutral in tone and possess a moderate level of subjectivity. Such reports are not necessarily very readable. This evidence is counterintuitive and contradicts previous findings on this topic.

Furthermore, the conformity score metric, along with the text data annotation approach, can be easily translated into editorial practical use and tailored for different conformity standards. The evaluative framework's emphasis on technical discussions about scientific aspects of reviewed research lays a theoretical foundation for peer-review report quality assessment of journal publishing policies. Additionally, the proposed text classification algorithms can be deployed as an integral part of a journal's manuscript processing system, contributing to accountable peer-review quality assurance against superficial or irresponsible comments. Moreover, the metric can also serve as an instrument that guides reviewers through the review processes, improving their manuscript comments (Rajesh et al., 2013). That instrument can be further utilized as a metric for reviewer performance measurement which is increasingly considered to be an useful evaluative indicator for a researcher's academic performance and should be an incentive for a researcher's academic career promotion endeavors (Májovský, 2025).

[5] Arous et al. (2021) refer to the review guidelines proposed by a number of journals, conferences and publishers and devise eight conformity criteria. These criteria address three aspects of review conformity: clarity, consistency and objectivity.

Literature review

Algorithm-based examination of peer review reports

The emergence of language modeling algorithms has enabled peer review report evaluation by means of algorithm programming. In this regard, some researchers have aimed to interpret the semantic meaning of peer review reports. Many of these attempts are exerted on individual, isolated sentences extracted from the review report text. For example, Bharti et al. (2022) proposed the Informativeness Score, which measures the extent to which a review report discusses the critical aspects (e.g., novelty, robustness) of a manuscript and its key sections. Using an annotated sentence dataset, Ghosal et al. (2022) trained a text classification model which classified review report sentences into multiple categories concerning different aspects of the review report discussion. Likewise, Okasa et al. (2025) performed text classification tasks to identify whether a review report sentence belonged to a corpus category indicative of a certain section (e.g., Methods). Moreover, Hren et al. (2022) used text content clustering to classify sentences from research grant evaluation reports into various themes addressing different aspects of grant proposals. By grouping research evaluation texts, the abovementioned research illustrates a means of inferring semantic meaning from review texts. It is possible that certain aspects of a reviewer's discussion are worth more attention than others; as such, focus should be placed on the core aspects of review report content when assessing for semantic meaning.

Some studies have examined peer review reports from the perspective of statistical features. Arous et al. (2021) employed review length, number of references, and keyword occurrence (e.g., "equation", "section", "figure") as model input features for predicting review conformity. Similarly, Meng (2023) predicted review quality based on a set of statistical features (e.g., review length, Dale–Chall readability score) as modeling predictors. Maddi and Miotti (2024) showed that review report length is associated with citations; however, this association is only significant for review reports that contain at least 947 words. Ausloos et al. (2016) examined the quantity, variety, and distribution of words in review reports. Referencing Zipf's law, this study evaluated review quality depending on the extent to which these statistical features deviated from those of common reports. In addition, Wolfram et al. (2021) examines review reports from the perspective of frequencies of hedging terms used and research-related terminologies. These authors attempted to gauge review report quality using a quantitative and aggregated measure that examined a review's language content as a whole, but scrutiny of a review's constitutive components (e.g., sentences) is necessary when carrying out a detailed examination for an in-depth understanding.

Some researchers studied the language contents of review reports from the perspective of the sentiment conveyed. Luo et al. (2021) used two dictionary-based sentiment analysis algorithms (TextBlob and VADER) to gauge reviewer sentiment and found that it is correlated with reviewers' willingness to accept the manuscript. Chakraborty et al. (2020) proposed aspect-based sentiment analysis, which binds sentiment polarity to a variety of aspects (e.g., appropriateness, clarity) of the reviewer's report. Their study found that sentiment aligns with reviewers' final recommendations. In Chakraborty et al. (2020)'s study, sentiment polarity is incorporated as a predictor of reviewers' recommendations. Bornmann et al. (2012) textually compared closed peer reviews (CPRs) and public peer reviews (PPRs), showing the former are more likely to contain positive emotion words. The implication is that reviewers providing CPRs are more concerned about the selection of

manuscripts for publication than reviewers providing PPRs. The studies mentioned above demonstrated that reviewers' sentiment is closely related to their recommendations; however, whether reviewer sentiment is indicative of review quality remains debatable.

Regarding the semantic content of review reports, some studies have attempted to specifically identify arguments that indicate a reviewer's opinions are well-supported by justifications included in their review report. As is shown in Guo et al. (2023), a trained argument mining system can automatically identify the extent to which a review presents substantiation. Based on argument mining, Hua et al. (2019) analyzed two aspects of reviewer arguments: argumentative propositions and their type (evaluations of the reviewed work or suggestions for improvement). Cheng et al. (2020) focused on interactive, argumentative discussions between reviewers and rebuttals, proposing a new argument pair extraction (APE) task to overhaul their interrelations. These studies explored review report contents, investigating how the level of substantiation of arguments can reveal review report quality. This focus on justification can be further extended to an examination of whether the identified arguments address certain aspects of the work reviewed, since even well-justified arguments can be considered superficial (e.g., typo errors) if they are not necessarily indicative of a manuscript's overall quality.

In summary, the prior literature examining review reports using algorithms indicated effective approaches to automatically inferring the semantic meanings from review reports. Regarding the editorial practices of peer review organizers, text classification techniques—the underlying methods used in most of the aforementioned studies—should be utilized in a way that highlights the most critical aspects of reviewer discussions. Moreover, when it comes to facilitating manual review quality investigations, an evaluative measure that dissects and considers the entire content of the review could be important for deepening human evaluators' understanding of the semantics of review reports.

Conformity of peer review reports

In order to ensure rigorous and productive peer review processes, previous studies documented various quality-related standards, focusing on conformity to these standards. Van Rooyen (2001) proposed a Review Quality Instrument (RQI) to gauge peer review report quality, examining a number of aspects of review content. Specifically, the RQI metric is composed of seven questions concerning the extent to which a review comments on the significance of the research reviewed (e.g., originality, importance of research questions), its specific contents (e.g., methods, results), the writing style, and whether the reviewer's opinions are sufficiently substantiated. Superchi et al. (2020) conducted a survey to collect participant opinions regarding important items for assessing review report quality. The finalized assessment checklist included 14 items focused on five aspects: study importance, methodological robustness, results discussion, reporting and transparency, and the nature of the comments. In Falkenberg and Soranno (2018)'s study, participants were required to complete reviews following a set of standards, such as summarizing manuscript topics and evaluating manuscript significance. Similarly, Sculley et al. (2018) advanced a rubric involving aspects such as discussions about empirical results, manuscript novelty, and statistical significance to facilitate review quality evaluation. Furthermore, Rajesh et al. (2013) introduced a structured review template that encourages reviewers to improve review quality by commenting on key sections in detail (e.g., Introduction, Methods).

In addition to general review reports, some studies focus on the quality of statistical reviews, which primarily discuss statistical methods or analyses. According to Henly et al.

(2010), statistical reviews are more likely identify critical flaws in the research reviewed, proving more useful for editors in making decisions and authors in improving their manuscripts. Moreover, Henly and Dougherty (2009) showed that statistical reviews are of a higher quality than regular reviews, and those that focus on technical aspects (e.g., design and methods) are better than the reviews that concentrate on background aspects (e.g., theories and literature reviews).

With the emergence of generative-AI techniques, a strand of literature focuses on assessing AI-generated review reports. Marrella et al. (2025) utilized ChatGPT to assess a number of manuscripts and generate review reports, showing that the ChatGPT-generated review reports outperform the human-written review reports in terms of the ARCADIA score review evaluation framework.[6] On the other hand, some studies documented that AI-generated reports do not perform satisfactorily. For example, Saad et al. (2024) found that the review comments generated by ChatGPT are inconsistent with its final decisions about accepting/rejecting reviewed research. Further, according to Zhu et al. (2025), Large-Language-Models (LLMs) can simulate academic critiques but the generated comments are only aligned with broad comments of the actual reviews, the LLMs' ability to generate in-depth comments that are learned from logics and scientific reasoning of reviewed research is limited.

Overall, studies that discuss review report quality generally rely on a metric that examines whether certain aspects have been sufficiently discussed. These metrics ensure that reviewer discussions are comprehensive and informative. However, it should also be acknowledged that some aspects that reviewers are expected to address are not equally important in terms of reflecting the soundness of the research reviewed. As shown in studies examining statistical reviews, technical discussions about aspects such as methods are correlated with higher review quality. Thereafter, when assessing review report quality, it could be worth prioritizing discussions of technical aspects. Moreover, the studies about AI-generated review reports also implies that an evaluative framework that rewards real scientific critiques which essentially are hard to be reproduceable by AI is needed, thus calling into question about whether the existing evaluative frameworks can be adaptive in face of the challenges of the generative AI techniques.

Methodology

Sample construction

As mentioned in Sect. "Algorithm-based examination of peer review reports", examining review reports at the sentence level enables investigators to overhaul the review content in detail; in this way, effective review report evaluation can be assured. As such, we investigated our sample of review reports at the sentence level first. For model training, we extracted 11,150 sentences from first-round peer review reports from 389 MDPI journals. The sentences are collected from the whole review reports and then classified into categories. This approach ensures the distribution of sentence frequency in each category to the largest extent approximates the category distribution of the review report sentence

[6] The ARCADIA score framework evaluates research from a number of aspects such as "Relevance", "Originality" and "Interpretation results" and so on.

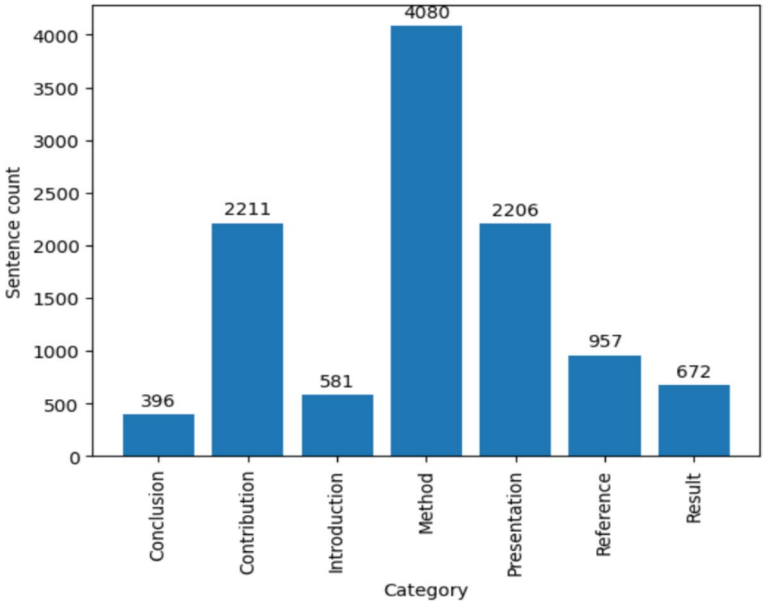


Fig. 1 The number of labeled sentences in each category (11,103 in total)

population. The classification result of a well-trained model should generate similar pattern of the category frequency with the population. We do not take a stratified sampling approach to avoid some sentences in the large categories are not used in model training. This is important in practice which requires continuously updating of the training dataset. In addition, we would also like the model to be trained to learn the sentence category distribution pattern of the training dataset. The publishing dates for these manuscripts were from 17 Mar 2010 to 7 December 2022, and the journals chosen covered multiple disciplines.[7] In that review report database, these journals have a total of 554,693 first-round review reports in which the review reports that contain the 11,150 sentences for model training are randomly selected at the journal level.[8] The review report numbers for each journal in the database are displayed in the Supplementary Materials. We manually labeled the individual extracted sentences according to the pre-defined categories, i.e., "Conclusion", "Contributions", "Introduction", "Methods", "Presentation", "References", "Results"; details are given in Sect. "Constructing the conformity score metric". When splitting a review text into a sentence corpus, we considered punctuation, such as periods, exclamation marks,

[7] The disciplines include: "Computer Science & Mathematics", "Engineering, Electrical & Electronics", "Society", "Business and Economics", "Physics", "Material Sciences", "Social Sciences & Psychology", "Environment & Ecology", "Engineering, Multidisciplinary", "Agriculture, Plants & Animal Sciences", "Clinical Medicine, Specific", "Chemistry", "Clinical Medicine, Multidisciplinary", "Biology & Biochemistry, Multidisciplinary", "Biology & Biochemistry, Biological Sciences", "Humanities", "Geosciences" and "Clinical Medicine, Healthcare".

[8] It is estimated that the 11,150 sentences are collected from 853 review reports. Since we do not record the report IDs associated with the sentences, we make the estimate based on the average number of sentences (13.01) of the review reports in the experiment dataset which is discussed in Sect. "The correlation between conformity scores and text quality indicators".

colons, semicolons, and questions marks, as the watersheds for text splitting because, generally speaking, these punctuation marks are followed by semantic shifts. In the setting of this study, therefore, in a broad sense, we called these split text corpora sentences. We excluded sentences containing "thank you", "congratulations", or "dear author" from the sample, as these are normally employed to express politeness and are irrelevant to discussions of the manuscript. We also did not take sentences with fewer than 5 words (5 exclusive) into account, as these sentences are unlikely to contain sufficient information. After pre-processing, the finalized model training corpus contained 11,103 sentences.

Figure 1 shows the distribution of the number of sentences contained in each class. The 4080 sentences classified as "Methods" account for the largest proportion of the training dataset. "Contributions" and "Presentation" are the second and third largest categories. The "Conclusion" category contains the smallest number of sentences. These findings indicate that reviewers tend to dedicate the most effort to discussing the research methodology; hence, methodological rigorousness is the most critical concern when assessing review quality.

After calibration on the abovementioned dataset, the model was applied to classify sentences from an experimental dataset containing text from first-round review reports of a number of MDPI publications from 2025. Using this dataset, we took all MDPI journals with 2025 publications into account. These journals follow the same generally consistent review quality assessment and conformity guidelines used by other peer review organizers. The text classification method proposed in this study can be tailored to suit probable review content guidelines. To determine the number of publications included in the dataset, for each journal, we rounded and randomly selected one-tenth of its total review reports which comment on publications in from 1 Jan 2025 to 11 Apr 2025. Regarding the sentence splitting method, we followed the same criteria as the abovementioned manual sentence-collecting process and completed the task using a Python program that automatically executes the criteria after text is inputted. In total, there are 235,211 sentences in the final dataset, which were collected from 18,014 review reports corresponding to 16,350 publications[9] from 415 MDPI journals[10] in 2025. Pulling text data from a wide range of journals and thus subject areas ensured that the model has good generalizability, with potential application to review reports on any subject.

Constructing the conformity score metric

Next, we examined sentence "conformity" with peer review quality standards. In most cases, peer review organizers have specific guidelines that define the extent to which a review report conforms to their expectations. These guidelines are mostly "content-based", as they examine reports from a semantic perspective. Since the review reports in this study were collected from MDPI journals and should thus conform to the company's standards, we referred to these guidelines[11] and classified sentences into seven categories depending on their semantic meaning. The seven categories are as follows: Methods, Results, Contributions, Conclusion, Introduction, Presentation, and References. To the best of our

[9] Review reports were randomly selected, and there was thus no guarantee that each publication included in the sample had every corresponding review report available.

[10] New MDPI journals have been launched since the sampling period for the training corpus.

[11] Those guidelines are a general guidance and every MDPI journal regardless of the differences in subjects follows those guidelines.

understanding, these categories also correspond to the individual requirements listed in the guidelines (Van Rooyen, 2001). In order to be classified into a category, a sentence must satisfy one or more of the requirements. The definitions of these categories are provided below.

Methods: Sentences classified in this category address reviewers' concerns about the methods employed while conducting the research. We believe that comments of this nature address the technical content of reviewed articles and should contribute significantly to improving the rigorousness of the reviewed research. In addition, discussing the methodological aspect of a study demands significant intellectual preparation and knowledge, signaling the reviewer's dedication to reviewing the research. We argue that reviewers intellectual preparation and knowledge are the dominating factors that add values to reviewed research and distinguish high quality review reports from common ones. This emphasis on methodological discussions in review quality assessment is also consistent with evidence showing that statistical reviews focusing on technical aspects tend to be higher quality than more general reviews (Henly & Dougherty, 2009; Henly et al., 2010). As such, we classify sentences related to this category into the top level of the conformity score hierarchy, assigning them a score of 3 in the scheme.

More specifically, we propose that the semantic meanings of sentences that are manually labelled as "Methods" should focus on the following aspects of the approaches employed in the reviewed research:

- Statistical modeling—discussions of statistical modeling involve the rigorousness of model set-up, the suitability of model use, limitations of the chosen models, and the potential of alternative models or parameters/variables employed in the model.
- Surveys/questionnaires—for reviewed research using surveys/questionnaires, a "Methods" comment discusses the justification of survey choices, survey purposes, survey types or participant selection.
- Data—research data are the basis for choosing a methodological approach; thus, reviewers' comments on research data are considered methodological discussions. Such comments address the motivations behind data choices, the justification of sample selection, the reliability of sources, and the preparation and processing of data. Furthermore, discussions of the limitations/potential flaws of data are also considered as part of the "Methods" classification.
- Research procedures—reviewer discussions of research procedures and analyses are also related to the methodology of the reviewed study. Such discussions might include an examination of the organization of the research steps, the reasonability of the research framework, the research procedures and their effectiveness/shortcomings.

Sentences within this category typically include words such as "method", "analyses","models" and "surveys/questionnaires".

Results—discussions regarding results are important for evaluating a study's significance. The implications of a study can be successfully upheld by results that feature convincing patterns. From the perspective of reviewers, commenting on a study's results and evaluating their value also demand extensive knowledge. However, the task of discussing results might be relatively easier than examining the methodology, as the comments do not rely as heavily on a thorough understanding of the technical details of the reviewed research. Unlike methodology-related reviewer comments which directly depend on reviewers' careful examination of technical details of reviewed research, comments about review research results are more likely to be superficial. Therefore, in our sample,

sentences that comment on the results of a reviewed article's analyses are in an independent category and ranked second in the scoring scheme (assigned a score of 2). In the process of manually labelling review sentences in this category, we include those that discuss analytical results, research findings or empirical evidence.

Contributions—an insightful review report should judge whether the reviewed article contributes to the existing literature and whether it has practical implications. Hence, we place such sentences second in the scoring scheme. Compared with other types of sentences, comments on a study's contribution have a wider scope, as strengths and weaknesses can revolve around a multitude of factors. We argue that discussions within this classification category should concern the following aspects of the research reviewed:

- Research novelty. This concerns whether the research reviewed is novel relative to the existing literature.
- Practical meanings. Discussions in this regard involve the practical values or regulatory implications of the research reviewed.
- Implications for future research. Discussions of this aspect indicates reviewers' opinions regarding whether the research reviewed illuminates directions for future research in the field.
- The importance of the subject matter. Reviewer discussions in this regard involve comments on the importance of the topic(s) or value of the research aims.
- Research originality (from an ethical standpoint). This aspect concerns the ethical consideration of whether the research reviewed can be seen elsewhere in the existing literature.

Such comments highlight the study's importance, which is the central concern regarding research quality for many review quality assessment rubrics. However, discussions in this regard are not as intellectually demanding, as a reviewer comments on novelty based only on prior knowledge (which could be subjective and limited) about developments in related research fields; in other words, they do not need to understand the technical details of the research reviewed to comment in this capacity. Answers to the question about whether research has a great contribution are easy to be general. As a result, we argue that sentences in this category fit into the second level of the scoring scheme, assigning them a score of 2.

Introductions—reviewers also tend to provide suggestions about a study's Introduction, in which the theoretical background and motivations are usually presented. Discussions of the Introduction often address the feasibility of the research setup, which is the basis of the article's contribution. Abstracts are, in a sense, similar to Introductions in that they provide a summative overview of the research's primary aspects. In our study, we combine reviewer discussions of these sections into the same category. Such comments rely on the reviewers' overall judgement about the value of the research reviewed, and a detailed understanding of how the analyses were carried out is not necessary (in most cases, such details are not given in an Introduction or Abstract). We therefore classify sentences in this category into the second level (score of 2) of the scoring scheme.

Conclusions—a study's conclusions provide a generalization of the overall content and evaluate whether the study's final verdict is reasonable. Similar to discussions about Introduction/Abstract sections, comments on Conclusions normally address the summative contents of the study and do not require a reviewer to have an in-depth understanding of the research reviewed. Sentences in this category are assigned to the second level of the scoring scheme (score of 2). In manual sentence classification, we include sentences

Table 1 The conformity score metric scoring design

Level	Category	Score
Level 1	Methods	3
Level 2	Results	2
	Contributions	2
	Introductions	2
	Conclusions	2
Level 3	Presentation	1
	References	1

This table displays the scoring scheme of the Conformity Score metric

that address conclusions concerning analyses, findings, or evidence in the "Conclusions" category.

Presentation—the quality of an article's presentation is often discussed by reviewers, as this affects readability. Presentation-related comments are generally concerned with organization or structure, the clarification of terms, the presence of a cohesive writing style, the clarity of figures/tables, and the accuracy of language. As such comments are not directly related to a study's semantic content, they are usually considered superficial, contain little intellectual information, and offer little in terms of expansion or improvement. A review report full of discussions of an article's presentation likely indicates that the reviewer did not dedicate themselves to exploring the reviewed research in detail. For this reason, such discussions are placed at the third conformity score level and are assigned a score of 1 in the scoring scheme.

References—some review reports also mention other studies in the existing literature, with the aim of enriching the article's analyses or discussions. We consider this an independent category, as such discussions indicate that a reviewer has used additional academic resources to support their review. However, in most cases, such comments do not require an in-depth understanding of the details of the research reviewed. A large number of reference recommendations do not reflect a reviewer's dedication to reviewing. Substantial intellectual preparation and knowledge are usually not needed for recommending related references. Therefore, comments in this category are assigned to the third level (scored 1) of the scoring scheme. Table S1 in the Supplementary Information shows some examples of the sentence classification.

The scoring scheme is summarized in Table 1.

In cases where a sentence straddled multiple categories, we classified it into the category on which it placed the most emphasis. For example, a sentence that discusses both the Methods and Contributions would be classified into the "Methods" category if words about modeling (or other aspects related to this category) were more frequent than words relating to research novelty (or other aspects related to the "Contribution" category). Moreover, we classified sentences that emphasized multiple sentence categories equally based on which category was discussed first. This is because a category that is discussed earlier on in a sentence is more likely to be considered important by a reviewer (e.g., the primacy effect[12]). Those sentences fit within multiple categories and were assigned a second or third label.

[12] The primacy effect describes the cognitive tendency that the first piece of information is more likely to be paid attention to than pieces of information in later positions (Van Erkel & Thijssen, 2016).

Sentence labelling was carried out by two researchers who have PhD degrees and are experienced in-house editors at MDPI. Both of them have experience in reviewing review reports of submissions to MDPI journals and have a solid understanding of the classification criteria mentioned above. The labelled data were further validated by their cross-checks which reach a consensus about the labelling of the data. Hence, the accuracy of the labelled data is, to a large extent, supported.

To facilitate the overall quality assessment of an individual review report, we propose the conformity score metric. This is an aggregate measure of a review's conformity with quality assessment guidelines, such as the RQI, depending on the individual sentence categories as abovementioned. Specifically, for a review report, we assign each classified sentence a score corresponding to the category score as described above, then we sum up the scores of all sentences in that review report to obtain the Conformity Score of that review report. The computation of the Conformity Score of an individual review report is built upon the following formula:

$$conformity_score_i = \sum_{j=1}^{7} category_score_j \times sentence_count_{i,j}$$

where $conformity_score_i$ is the conformity score of review report i; $class_score_j$ is the score assigned to the category j; and $sentence_count_{i,j}$ is the number of sentences classified into category j for report i.

The LSTM model

To perform the multi-class text classification task, we used the LSTM model, which proved successful and computationally efficient in dealing with language data. As mentioned above, we established a corpus containing 11,103 labelled and pre-processed sentences. We randomly split the corpus database into training (90%) and testing datasets (10%), comprising 9992 and 1111 sentences, respectively. Figure 1 displays the distribution of the sentences from each category in the training dataset.

Before applying the model, we removed the "stopwords" and numbers from the sentence text corpus, as these do not hold significant semantic meaning. After that, we created a dictionary of tokenized words in the sentence text corpus and assigned each word a unique index. The dictionary includes the most frequent 10,000 words tokenized from the corpus and excludes any punctuation. After processing, we generated a sequence of encoded words for each sentence in the corpus. The sequences were then padded to equal lengths of 150 elements, with the encoded words being placed last and the other early positions being filled with 0.

We applied the LSTM model to the testing dataset and achieved 95% accuracy by accounting for alternative labels. To cross-validate the effectiveness of the model, we evenly split the whole shuffled corpus into 10 folds with each fold containing 1110 sentences. We use each of the folds as the testing sample and apply the trained model to predict the classes of the sentences in the testing sample. The average accuracy rate of these 10 tests is 96% taking into account alternative labels.

Next, we fed the processed sequences into the embedding layer of the LSTM model. The embedding layer is the first layer, and in addition to assigning 10,000 input dimensions and 150 as the input sequence length (in alignment with the parameters of the abovementioned encoded word sequences), the input dimensions in this layer were represented by

128 dimensional. In the second layer, we used spatial dropout and set the dropout rate to 0.7. This layer was connected to a LSTM layer with 64 units or neurons, a 0.7 dropout rate and a 0.7 recurrent dropout rate. The output layer is a fully connected layer that has 7 units and uses "softmax" as the activation function. In terms of model fitting, we used the Adam optimizer, the categorical cross-entropy loss function and accuracy as the evaluation metrics. We used grid search to ensure that the chosen parameters were optimal.

Statistical features of review reports

We performed extensive data preprocessing and feature engineering to prepare the dataset for analysis. Meng (2023) produced a set of text quality indicators that they argued are predictive of review conformity. In this study, we analyzed the correlation between these indicators of review quality and conformity scores based on the experimental dataset. The data preprocessing steps and the feature extraction methodologies applied are described in detail below.

Linguistic features

The raw review report text data were in HTML format. After we loaded the peer review data, we utilized BeautifulSoup, a Python library for processing HTML/XML files, to strip HTML tags from the text to provide clean and readable text. We then computed the following linguistic features of the individual review reports in the experimental dataset:

Recommendation Score (rec_score)—to quantify the reviewer recommendations, numerical scores are assigned to different recommendations: 1 for "accept", 0 for "major/minor", and −1 for "reject".

Review Length, Original and Log-transformed (rev_len_log)—the length of the review is calculated in terms of word count, and a log transformation is applied to handle skewness. For each review report HTML/XML file, we count the number of tokenized words after punctuation marks are removed.

Average Sentence Length (avg_sent_len)—this feature is computed by dividing the total number of tokenized words by the number of tokenized sentences.[13]

Number of Unique Words (num_unique)—the number of unique words (used to assess the lexical diversity of the peer review).

Number of Sentences (sent_num)—the total number of tokenized sentences in each review.

Number of Reference Words (num_ref)—the frequency of the reference words, including "section", "page", "figure", "paragraph", "p.", "fig." "equation," and "line".

Dale–Chall Readability Score (readability)—to evaluate the readability of the text, we employ the Dale–Chall Readability Score (Dale & Chall, 1948), which is used to examine the level of reading difficulty for readers with various vocabulary levels, using the Python package textstat. This score is also widely recognized as a metric in the field of readability assessment and provides an estimate of the text's complexity based on familiar word usage and sentence structure. The text readability score is determined based on the number

[13] We used the NLTK library in Python to split sentences. The NLTK library uses pre-trained models to identify individual sentences in a text. The "Number of Sentences" feature described below also depends on the same tokenization method.

Table 2 Summary statistics of the classified texts of the review reports in the experimental dataset

	Mean	Median	Max	Min
Conclusions	1.2609	1.0000	10.0000	1.0000
Contributions	2.6792	2.0000	38.0000	1.0000
Introduction	1.4035	1.0000	7.0000	1.0000
Method	7.3107	6.0000	124.0000	1.0000
Presentation	2.4950	2.0000	34.0000	1.0000
References	1.6823	1.0000	15.0000	1.0000
Results	1.7790	1.0000	15.0000	1.0000

This table displays the frequency of the sentence classes based on 18,014 *IJMS* review reports. The sentences are classified by the trained LSTM model according to the Conformity Score metric

of difficult words and the average sentence length, with more complex sentences being assigned a higher score.

Sentiment features

Subjectivity (sub)—TextBlob's sentiment analyzer was used to determine the subjectivity of a review text.

Polarity (pol)—TextBlob also provided the polarity scores, indicating the overall sentiment of a text (e.g., positive/negative).

The sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [−1.0, 1.0], which indicates sentiment varying from negative to positive. The subjectivity is a float within the range [0.0, 1.0], where 0.0 is very objective and 1.0 is very subjective.

Interactive Features

Review–Abstract similarity (rev_abst_simi_specter)—to measure the content overlap between reviewers' comments and the corresponding abstracts, we employed the SPECTER model (Cohan et al., 2020), a transformer-based architecture specifically designed to assess the similarity of scientific documents. This model allows documents to be embedded into a dense vector space where semantically similar documents are close to one another.

Results and discussion

The correlation between conformity scores and text quality indicators

According to Meng (2023), a review report's text quality indicators are predictive of its conformity with assessment guidelines. In this study, we assume that the proposed Conformity Score indicates the extent to which the contents of a review report conform with certain assessment guidelines. Hence, the text quality indicators, as described in Meng (2023), should be correlated with Conformity Scores if said scores are reflective of the quality of a review report. We examine the statistical text quality features, as discussed

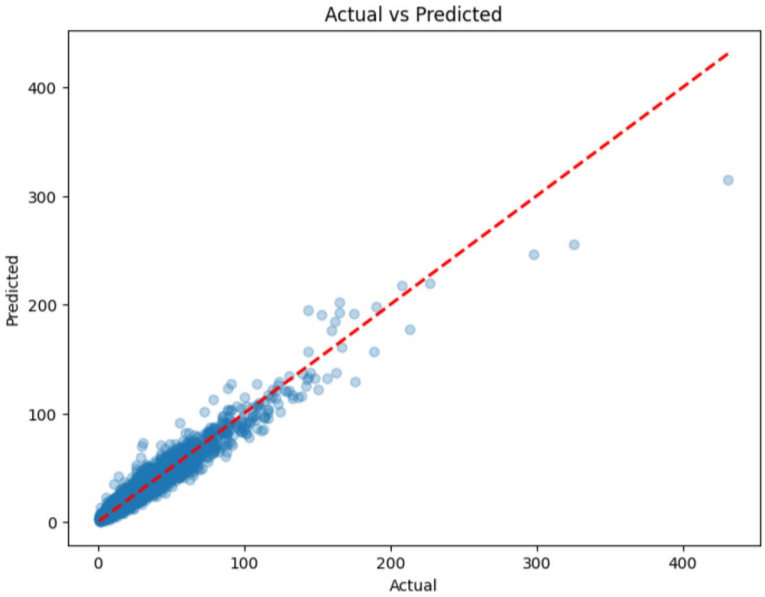


Fig. 2 Actual observations of predicted and actual Conformity Scores. This figure displays the actual observations of the actual Conformity Scores as well as those predicted by the XGBoost model. The input features of the model are the statistical features of the review report texts, as defined in Sect. "Statistical features of review reports". The sample comprises *18,014* review reports

in Sect. "Statistical features of review reports". The examination is based on the 18,014 review reports in the experimental dataset as mentioned in Sect. "Sample construction". We also apply the trained model, as discussed in Sect. "The LSTM model", to classify the individual sentences within the reports, and we compute the Conformity Scores based on the sentence classification and the scoring metric. Table 2 shows the frequency of the classes of the total 234,345 sentences of those review reports. It can be learned from Table 2 that each class has sufficient data representation since none of the means of sentence frequency of each class is extremely marginal, and the variations in the means across the classes is very consistent with the sentence frequency distribution of the model training sample (Fig. 1), we thereby argue that the trained model applied to the experimental dataset captures the characteristics of the sentence frequency distribution of the classes in the model training sample. Specifically, discussions related to the research methods have the highest frequency, with a mean of 7.3107, indicating that reviewers generally focus on technical solutions to research questions within the reviewed submissions. The Contributions and Presentation categories have average frequencies of 2.6792 and 2.4950, respectively, indicating that reviewers are also concerned about each study's novelty/strengths/weaknesses and readability. Very few comments relate to the Introduction and Conclusion, possibly because these sections contain little content in need of detailed reflection.

We argue that the set of statistical features of those review reports in the experimental dataset should be correlated with the Conformity Scores of those review reports, since the statistical features jointly reflect review report quality. Therefore, we employ a machine learning model XGBoost which has the statistical features of the review reports as inputs to predict Conformity Scores of the review reports. After the grid search identifies the best parameter combination, we assign the model with 200 trees in the ensemble, 70% of

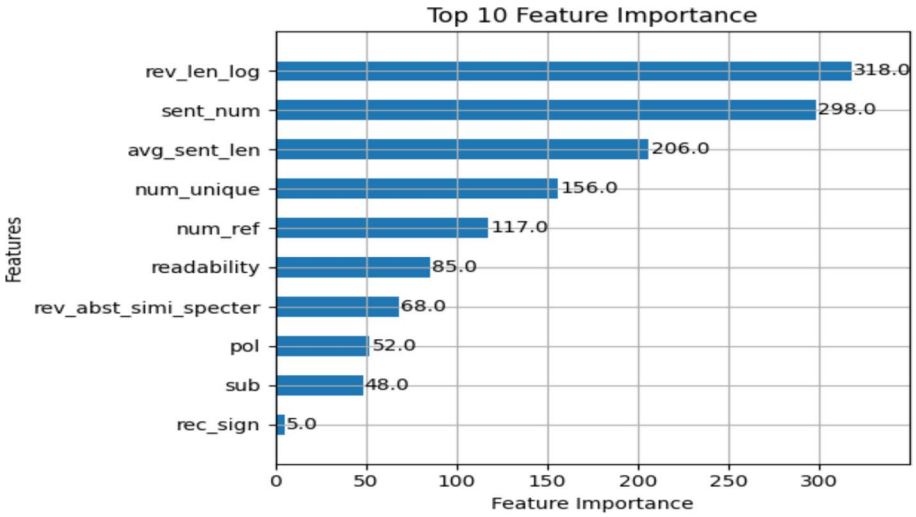


Fig. 3 Feature importance of the XGBoost model. See Sect. "Statistical features of review reports" for definitions

randomly sampled features for each tree, 0.1 learning rate, 3 max depth of each tree, 1 minimum sum of instance weight needed in a child, 90% of randomly sampled training data.

Figure 2 shows the prediction result. The predicted Conformity Scores vary linearly with the actual results, and the model achieves an R squared value of 0.93. Hence, the model that utilizes the text quality indicators as predicting variables effectively predicts the Conformity Score of a review report. We confirm that the Conformity Score metric is indicative of review report quality.

According to Meng (2023), the variables in the feature set in a sense indicate overall review report quality and are strongly correlated with the conformity standards of publishers, journals or conferences, all of whom are highly concerned about the quality of collected review reports. In this study, we assume that the proposed Conformity Score metrics reflect some aspects of the conformity standards described by Meng (2023). For example, we argue that mentioning different sections of a reviewed study justifies a reviewer's decision; this aligns with the "Justifications" standard discussed by Meng (2023). In addition, we highlight the importance of methodology discussions, which are indicative of reviewers' understanding of the technical details of reviewed submissions, while we give little weight to discussions about articles' presentation, as these tend to be superficial. These arrangements are consistent with the "Objectivity" standard outlined by Meng (2023). Furthermore, the Conformity Score metrics are developed based on MDPI's guidelines for their in-house editors' review report-evaluating practices; hence, the Conformity Score metric is representative of the standards that MDPI expects of its reviewers. Therefore, the Conformity Score metrics generally reflects the expectations of review reports within academic publishing.

As shown in Fig. 2, the statistical features of the review reports are correlated with the Conformity Scores, which separate the textual content of the review reports and indicate its compliance with MDPI standards. The previous literature has demonstrated that those statistical features are indicative of quality of review reports; thus, the correlated pattern in Fig. 2 implies that the Conformity Score metrics can be an effective indicator of the quality

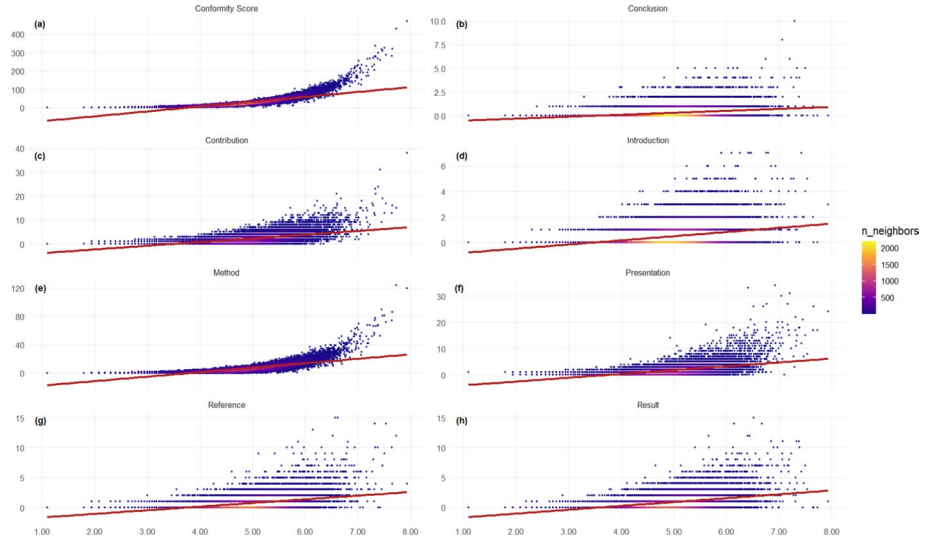


Fig. 4 The relationship between review report length (rev_len_log) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with review report length. Figs **b** to **h** show how the frequency of a sentence class varies with review report length. The trend is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

of review reports. The correlated pattern also indicates that a review report's conformity with content-based quality standards can be quantified and the quantitative measurement is interrelated with the statistical quality features. Therefore, it is conceivable that human editors' evaluation of review reports with reference to peer review organizers' conformity standards can be facilitated by an automatic algorithm-based program that parses the semantic meaning of review report content.

Figure 3 shows the statistical importance of the features used to predict the Conformity Scores of the review reports in the experimental dataset. The features rev_len_log, num_unique, avg_sent_len and num_ref have relatively high statistical importance when it comes to effectively predicting Conformity Scores. Those features are strongly correlated with review report length. Also, sent_num, which measures the number of sentences in a review report, has very strong relation to variation in Conformity Scores. This is not surprising, as Conformity Scores are an aggregate measure of sentence counts. In addition, as indicated by its feature importance, a review report's readability also contributes to the machine learning model's prediction of Conformity Scores, accounting for the interpretation that review reports containing a large number of technical discussions have low readability. Compared with the abovementioned features, the feature of review–abstract similarity (proxied by rev_abst_simi_specter) has relatively low statistical importance when it comes to predicting Conformity Scores, but that feature importance is still stronger than that of the two sentiment-related measures, sub and pol. Hence, it is likely that sentiment indicators have little relation to review report quality in terms of conformity. rec_sign ranks lowest among the features, implying that reviewers' recommendation is not relevant to the review reports' conformity. Overall, Fig. 3 implies that the features relating to content

length are the most effective predictors in the machine learning model, while sentiment measures and reviewers' recommendations are of least importance.

Review report length and conformity scores

Next, we examine how each individual statistical feature varies with the Conformity Scores. Figure 4 shows that the Conformity Scores are positively correlated with rev_len_log, indicating that review reports of comprehensive contents are likely to have higher Conformity Scores. Length is an important indicator of review report quality, and it is understandable that the authors of more substantial reports provide more constructive comments to justify their recommendations, prompting a high Conformity Score. Also Maddi and Miotti (2024) indicates that long reports require substantial author revisions and thus contribute to enhancing quality and visibility of reviewed research. This is consistent with the findings of (Buljan et al., 2020), who show that the length of review reports is also related to reviewers' recommendations (e.g., recommendations of major revisions are associated with longer review reports compared with recommendations of acceptance, which do not require substantial justification).

The individual sentence classes also have a positive correlation with review report length, and the "Method" class has the strongest correlation. This is not surprising, as discussions about the methods provide the greatest contributions to review reports (Table 2). It can be inferred that technical discussions, which we believe are a determinant of the value of a review report, vary with review report length, which is generally considered as an indicator of quality. More than that, We report the R-square values of the approximated linear relationships of all the sub-figures in Table 3.

Average sentence numbers and the conformity scores

Figure 5 displays the relationship between review reports' average sentence number and their Conformity Scores, which is generally linear, as shown in Fig. 5. The Conformity Score metric, as an aggregate measure of sentence count, in nature favors the review reports of a large number of sentences. The interpretation can be that a great number of sentences imply a high standard of review report quality. Additionally, the "Method" class has the strongest correlation with average sentence numbers (Table 3). Therefore, the reviewers who make greater efforts to address the scientific aspects of reviewed manuscripts tend to use a large number of sentences. Splitting discussions into multiple sentences in a review report not only facilitates authors' reading and understanding, but also allows them to address various aspects of the reviewed articles in greater detail. Interestingly, the "Presentation" class also has a positive correlation with average sentence numbers. Hence, the reviewers who could be not very dedicated also favor using more sentences to describe their concerns about articles' presentation.

Average sentence length and the conformity scores

Figure 6 shows that the Conformity Score metric and the individual sentence classes do not have a clear linear relation with the average sentence length of the sample review reports. In addition, it can be inferred from Fig. 6 that the review reports that have a high Conformity Score and a large number of discussions on the methods of reviewed manuscripts are generally composed of shorter sentences (the highest conformity scores roughly correspond

Table 3 R-square of the statistical features

	Conformity Score	Conclusions	Contributions	Introduction	Method	Presentation	Reference	Results
rev_len_log	0.652***	0.071***	0.320***	0.101***	0.568***	0.269***	0.175***	0.168***
avg_sent_len	0.002***	0.000**	0.002**	0.000	0.002***	0.001***	0.000	0.000
sent_num	0.714***	0.064***	0.319***	0.084***	0.604***	0.395***	0.216***	0.171***
num_unique	0.422***	0.025***	0.171***	0.051***	0.375***	0.251***	0.135***	0.096***
pol	0.014***	0.000	0.001***	0.000**	0.016***	0.009***	0.002***	0.001***
sub	0.000**	0.000**	0.001**	0.000	0.000**	0.000***	0.000*	0.000***
rev_abst_simi_specter	0.082***	0.005***	0.052***	0.012***	0.077***	0.009***	0.014***	0.013***
readability	0.042***	0.009***	0.025***	0.006***	0.029***	0.049***	0.013***	0.024***
num_ref	0.000	0.000	0.003***	0.001***	0.001***	0.088***	0.000	0.000

This table shows the R-square of the statistical features. We apply the univariate regression model, which has an individual statistical feature as the independent variable and a Conformity measurement as the dependent variable. The sample comprises 18,014 review reports

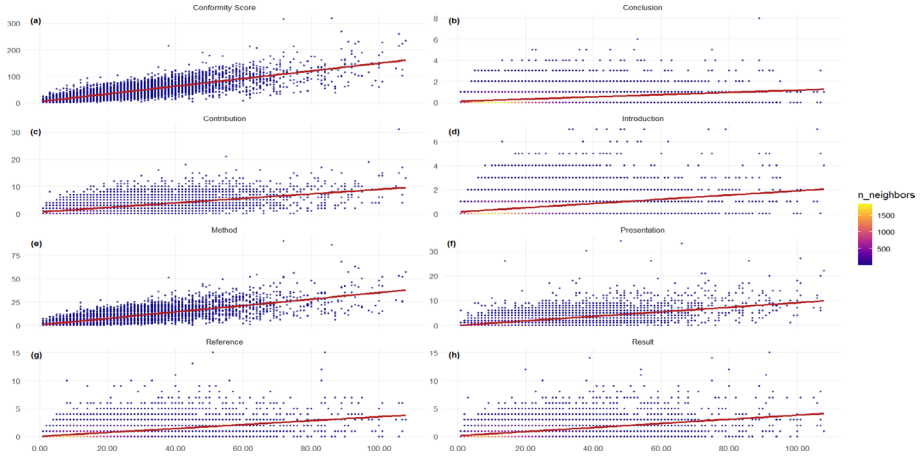


Fig. 5 Average sentence numbers (x-axis) and conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with average sentence numbers, while Figs **b** to **h** show how the frequency of a sentence class varies with average sentence numbers. The trend is fitted by a linear regression model. The sample comprises 18,014 review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

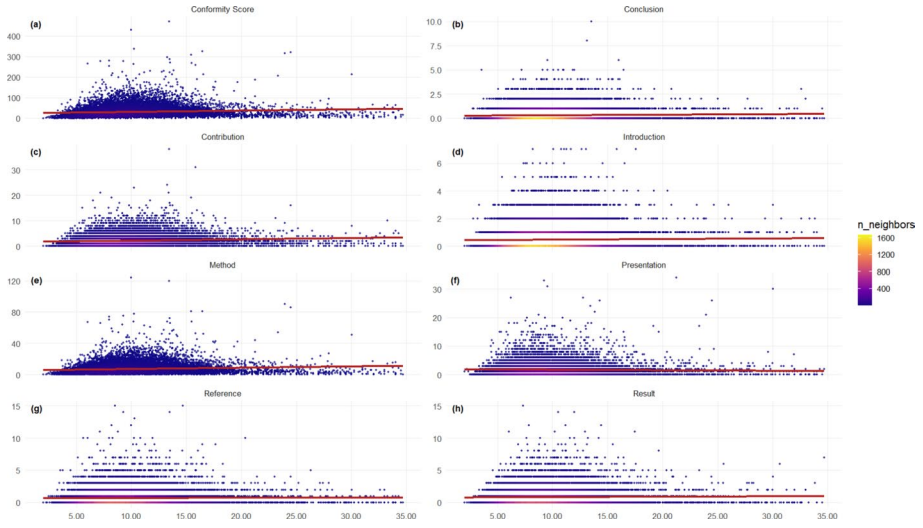


Fig. 6 This figure shows the relationship between average sentence length (avg_sent_len) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with average sentence length. Figs **b** to **h** show how the frequency of a sentence class varies with average sentence length. The trend line is fitted by a linear regression model. The sample comprises 18,014 review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

to 10 tokenized words). This implies that discussions about the technical aspects of a study do not necessarily need to include long sentences; a large number of short and clear

sentences could include more details. However, some reviewers tend to write long, unbroken sentences, leading to a low Conformity Score and a low number of method discussions. Although such sentences are constructive in essence, consolidating several aspects of a method into a single sentence impairs authors' understanding of the critique. Therefore, the Conformity Score metric, which relies on sentence counts rather than sentence lengths, could be seen to penalize more complex sentences that address multiple concepts.

Number of unique words and the conformity scores

Similarly, the number of unique words included in review texts (num_unique) is positively correlated with Conformity Scores (Fig. 7), indicating the comprehensiveness and informativeness of a review, as multifaceted discussions generally require the use of a large number of different words. According to Ausloos et al. (2016), a rich vocabulary facilitates authors' understanding of the reviewers' suggestions and helps to prevent redundant statements. In addition, using interchangeable words to express the same meaning can improve the readability of review texts. This may also explain why high-quality review reports contain many unique words. Additionally, "Method" is the individual class with the strongest correlation with the number of unique words. This is understandable, as detailed discussions about a study's research methods of a study necessitate the use of a range of words.

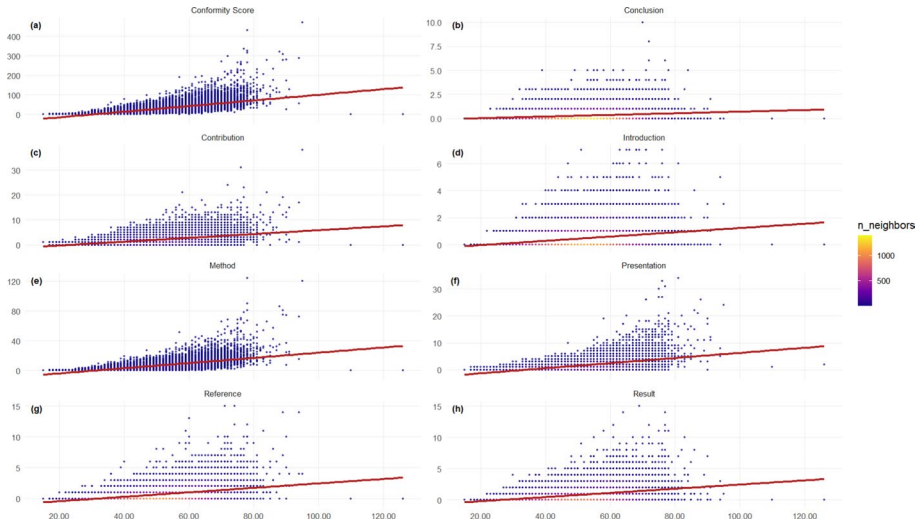


Fig. 7 This figure shows the relationship between the number of unique words (num_unique) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with number of unique words. Figs **b** to **h** show how the frequency of a sentence class varies with the number of unique words. The trend line is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

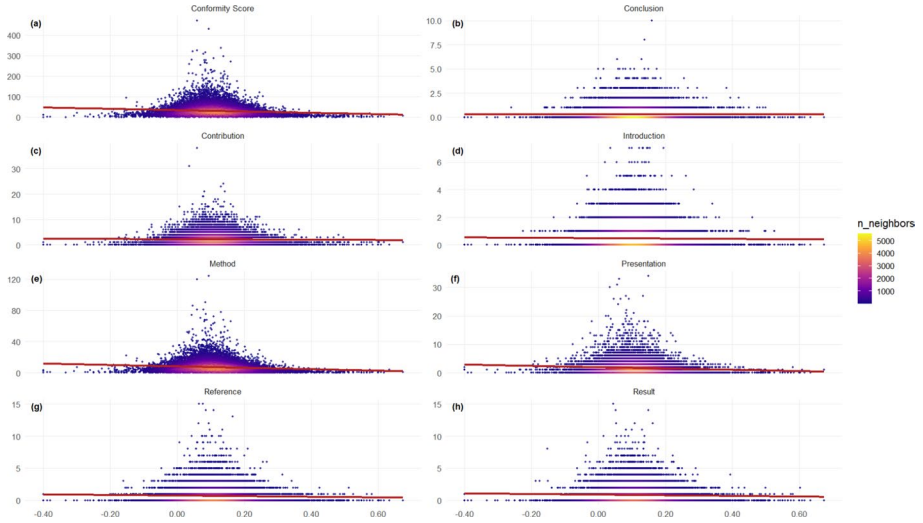


Fig. 8 The relationship between sentiment (pol) and the conformity measurements. Fig **a** shows how Conformity Scores (y-axis) vary with sentiment (x-axis). Figs **b** to **h** show how the frequency of a sentence class varies with sentiment. The trend is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

Sentiment and the conformity scores

A number of previous studies have examined the sentiments conveyed by reviewers (Meng, 2023; Sizo et al., 2019; Wang & Wan, 2018). Meng (2023) found that high-quality review reports usually express neutral sentiments as opposed to overly positive or negative views. Also, as indicated by MDPI's review evaluation guidelines, reviewers' views should be presented politely. Henceforth, neutrality conforms with the expected quality standards for review reports. Figure 8 shows that neutral review reports[14] have high Conformity Scores, in agreement with the findings of Meng (2023). However, it could also be suggested that review reports expressing negative sentiments are generally of high quality, as they demonstrate the reviewers' willingness to provide critical comments. The evidence presented in Fig. 8 violates this view and indicates that high-quality review reports are not necessarily the strictest; rather, objectivity is important when writing a meaningful review. Moreover, some studies have found that the sentiment of review texts predicts reviewers' recommendations (e.g., a negative sentiment leads to a "reject" recommendation) (Ghosal et al., 2019; Wang & Wan, 2018). The highly qualified review reports could generally be linked to revision recommendations (e.g., "major" or "minor" recommendation), which determine the acceptance of a reviewed article based on the results of revisions. With regard to the individual sentence classes, they show very similar patterns to Figure (a), whose distribution concentrates on the neutral sentiment level. This evidence implies that reviewers

[14] The sentiment polarity values in the center of the distribution are slightly greater than those in the neutral polarity watershed (0), but these values are still considered to be neutral in a broad sense.

expressing neutral sentiment tend to provide more in-depth discussion than those whose sentiments are more extreme.

Subjectivity and conformity scores

Figure 9 shows that review reports with high Conformity Scores are moderately subjective (sub). Yuan et al. (2022) mention that a review is typically concerned with both objective aspects (e.g., the factual issues of a reviewed article, such as insufficient details in methodology illustrations), and subjective aspects (e.g., opinions about the contributions of a reviewed article). Park et al. (2014) found that reviewers' subjectivity benefits scientific communities, as subjective comments allow reviewers to express their opinions without the influence of peer pressure. Review reports scored highly by the Conformity Score metrics could include both personal feelings regarding the issues within a reviewed article and objective descriptions of established facts. This explains why high Conformity Scores correspond to a moderate level of subjectivity. The individual sentence classes show similar patterns to Fig (a).

Review abstract similarity and conformity scores

Relevance to the content of reviewed articles was an important consideration of prior studies examining review report quality (Falkenberg & Soranno, 2018; Meng, 2023; Priatna et al., 2017). Figure 10 shows that rev_abst_simi increases with the Conformity Score. Review reports highly ranked by the Conformity Score metrics tend to have strong relevance to reviewed articles, and reviewers whose reports obtain a high Conformity Score

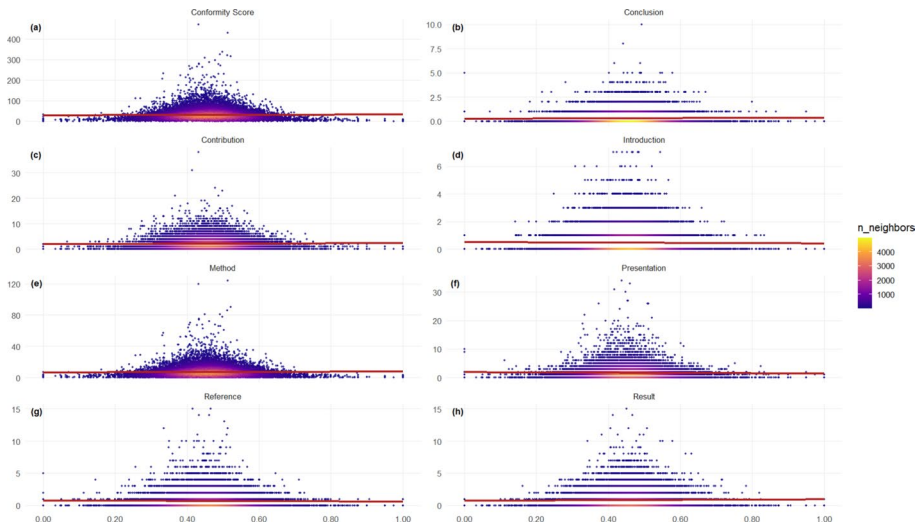


Fig. 9 This figure shows the relationship between subjectivity (sub) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with subjectivity. Figs **b** to **h** show how the frequency of a sentence class varies with subjectivity. The trend is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

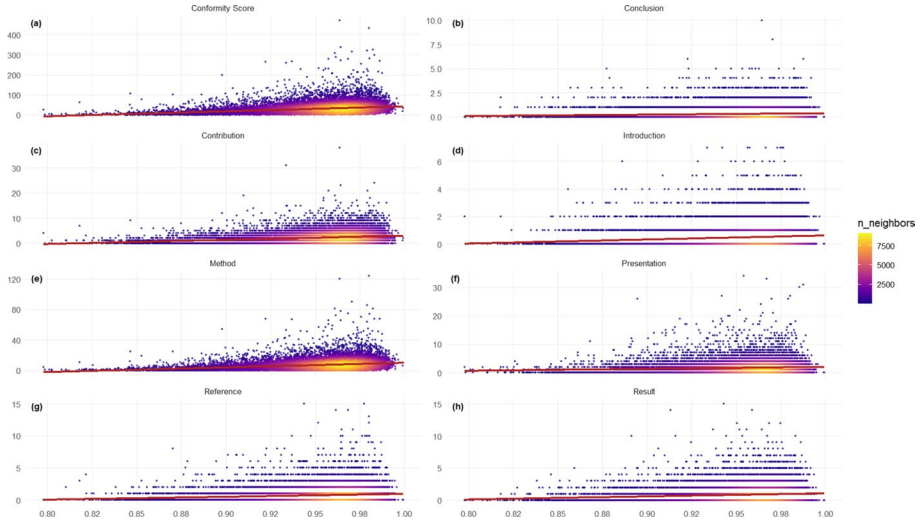


Fig. 10 This figure shows the relationship between review–abstract similarity (rev_abst_simi_specter) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with review–abstract similarity. Figs **b** to **h** show how the frequency of a sentence class varies with review–abstract similarity. The trend is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

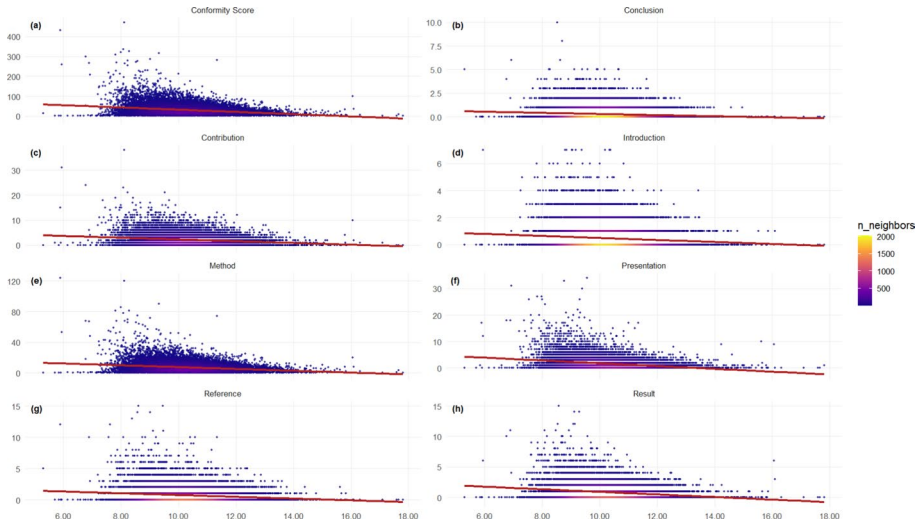


Fig. 11 This figure shows the relationship between readability (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with readability. Figs **b** to **h** show how the frequency of a sentence class varies with readability. The trend is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

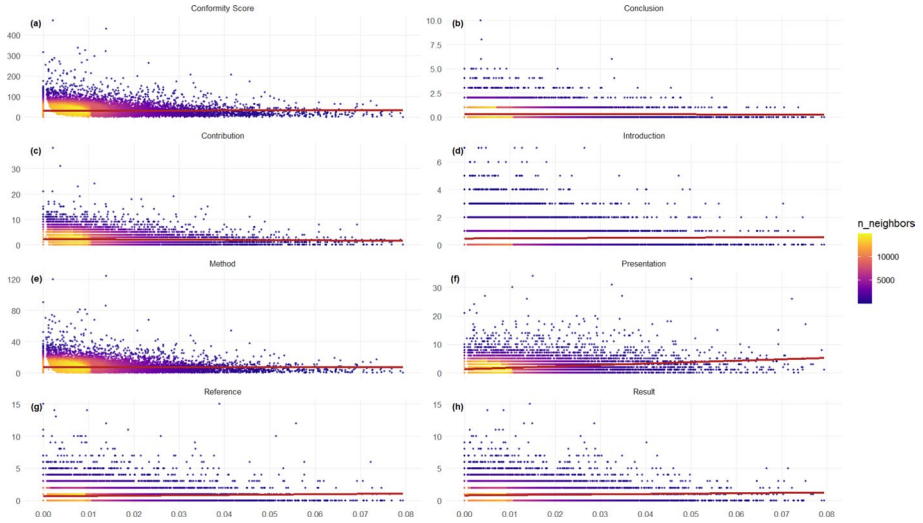


Fig. 12 Number of reference words and conformity measurements. This figure shows the relation between number of reference words (num_ref) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with number of reference words (num_ref). Figs **b** to (h) show how the frequency of a sentence class varies with the number of reference words (num_ref). The trend line is fitted by a linear regression model. The sample comprises *18,014* review reports. For displaying, we remove the outliers beyond the 5 percentile and the 95 percentile of the data points on the x-axis. The red line fits the linear relation between datapoint on the x-axis and the y-axis. The color bar measures the density of the data point distribution

likely develop their review reports with consideration of the reviewed articles' principal content. As indicated by Falkenberg and Soranno (2018), high-quality review reports contain a large number of comments on the articles' analyses and importance. In this sense, the pattern shown in Fig. 10 is consistent with the evidence of Falkenberg and Soranno (2018), as comments related to articles analyses and importance are highly weighted in the Conformity Score metrics. Among the individual sentence classes, the "Method" class has the strongest correlation with review–abstract similarity, likely because authors tend to outline their research methods in the Abstract. The "Contribution" class also has a relatively high correlation with review–abstract similarity; it is likely that the highlights of a study's strengths or implications within the abstract are often discussed by reviewers.

Readability and conformity scores

Figure 11 shows the relationship between readability and the Conformity Scores. The pattern shown in this figure is, in a way, counterintuitive, as review reports with high Conformity Scores generally have low readability. However, low-quality review reports that have low Conformity Scores usually contain superficial comments which are expressed in simple language, while high-quality reviewers utilize more complicated words and phrases to comprehensively express their thoughts. This could explain why the review reports with high Conformity Scores are generally of low readability, contradicting the assertion that high readability is a determinant of quality.

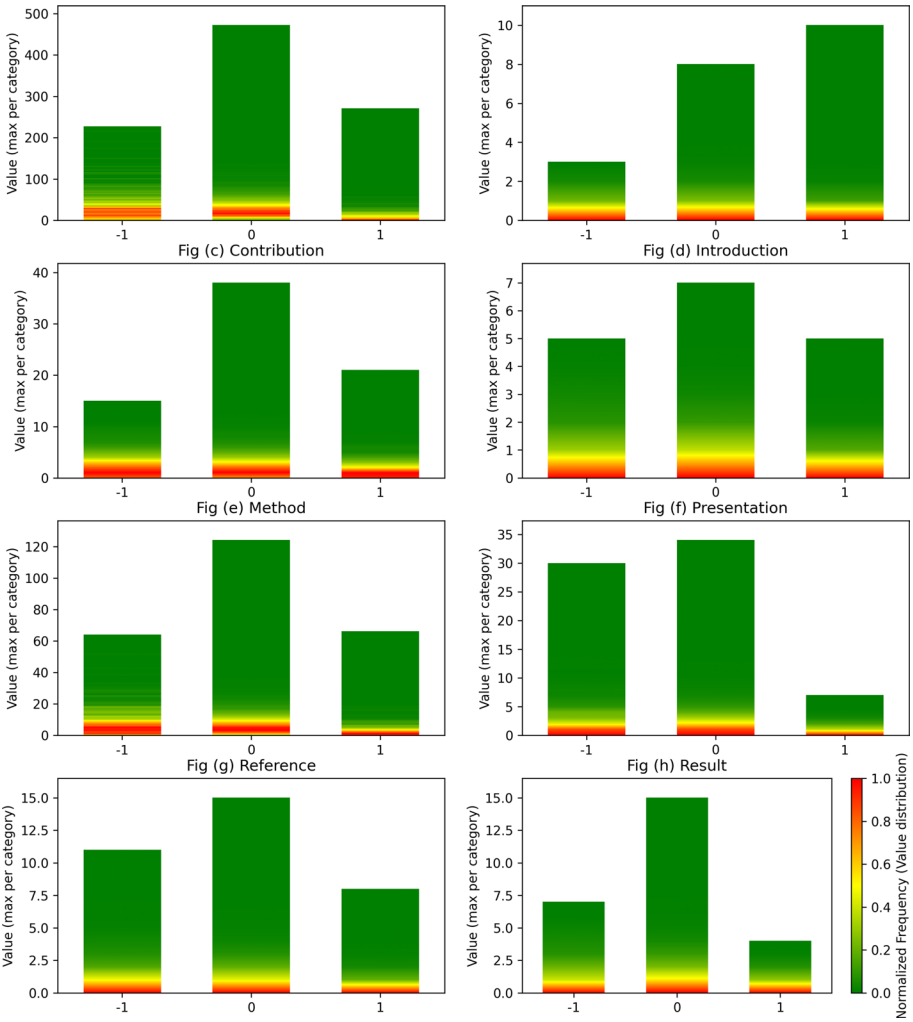


Fig. 13 This figure shows the relationship between recommendation signs (rec_sign) (x-axis) and the conformity measurements (y-axis). Fig **a** shows how Conformity Scores vary with the number of recommendation scores (recommendation_score). Figs **b** to **h** show how the frequency of a sentence class varies with the number of recommendation scores (recommendation_score). The trend is fitted by a linear regression model. The sample comprises 18,014 *IJMS* review reports. The color bar measures the density of the data point distribution

Number of reference words and the conformity scores

In a review report, reference words such as "section", "page", "figure", "paragraph", "p.", "fig." "equation" and "line" delineate a reviewer's comments on specific components of reviewed research. Fig (a) of Fig. 12 shows that the review reports with high Conformity Scores tend to mention reference words less frequently. It is possible that reviewers concerned with the technical content of reviewed research do not consider those specific components to be important. In addition, discussions about those specific details are usually

concerned with the layout or presentation of the reviewed research; thus, those discussions could be superficial and fail to touch on scientific aspects of the reviewed research. Moreover, Table 3 shows that has the sub-figure regarding "Presentation" has the largest R-squared compared with the other sub-figures, indicating the relatively strong relationship between the number of reference words and frequencies of the sentences classified as "Presentation". Overall, the patterns presented in Fig. 12 indicate that the reference words do not account for reviewers' commitment to technical discussions.

Recommendation scores and conformity scores

Recommendation scores imply the extent to which the reviewed research is of publishable standard. Fig (a) of Fig. 13 shows that "major/minor revision" recommendations tend to have higher Conformity Scores than other kinds of recommendations. It is likely that the review reports recommending further revisions contain in-depth discussions that the reviewers expect the authors to address. The scores for "acceptance" recommendations are slightly greater than the review reports that recommend "rejection". This evidence is consistent with the evidence discussed in Sect. "Sentiment and the conformity scores", implying that high standard review reports are not necessarily relating to negative reviewer opinions. Furthermore, Fig (f) and Fig (g) show that, compared with the "acceptance" recommendations, the "rejection" recommendations have higher sentence frequencies concerning the two lowest scored categories "Presentation" and "Reference" in the Conformity Score metric, inducing concern about irresponsible rejections of reviewed research.

Falkenberg and Soranno (2018) identify specific terms to investigate a report's relevance to the reviewed article. Those specific terms are either "evaluated entities", which revolve around entity classes such as the "analyses" or "goals" of a reviewed article, or "evaluative adjectives", which express reviewers' judgments (e.g., views on the comprehensibility or suitability) about a reviewed article. According to Falkenberg and Soranno (2018), highly relevant review reports contain a large number of comments in the "Analysis" class. Likewise, Priatna et al. (2017) assigned the highest weight to discussions about the methodologies of reviewed articles in the proposed scoring system.

Because review reports of academic articles are usually highly specialized and involve subject-specific terminology, assessing review texts may be difficult (Meng, 2023). This study shows that automatically examining the contents of review texts is feasible, as sentences can still be classified despite the presence of complex terminology. This is because the model is trained in a text dataset full of academic words whose features are clear in the context of the dataset.

Previous studies have attempted to automatically evaluate review report qualities (Arous et al., 2021; Meng, 2023; Priatna et al., 2017). Unlike the methods used in other studies, the Conformity Score metric does not rely on the structural features of review texts. In our view, the structural features of a review report comprise certain rules (e.g., the inclusion of specific words) which may cause the semantic meaning of a text to be misinterpreted. Therefore, we propose using a supervised machine learning method which enables the system to infer the semantic meanings of review texts. We argue that review report quality can be directly reflected by the semantic meanings conveyed in a report. In addition, the Conformity Score metrics assesses each sentence individually rather than taking features of a whole review report as input for a quality prediction model. This approach should provide a detailed examination of a review report and facilitate reliable quality prediction reflected by the finalized Conformity Score, which is a sum of contributions from each sentence.

Yuan et al. (2022) examine the possibility of generating AI-based reports and conclude that AI-generated reviews are "less constructive" and "less factual" than human-written texts; however, AI texts perform well when it comes to generalizing the key research ideas of scientific articles. In editorial practice, effectively identifying AI-generated review reports is a priority due to ethical concerns. In our view, detailed discussion of an article's methods is a very important feature that distinguishes human-written reviews from those that have been automatically generated, since it is highly challenging for an AI model to infer the probable flaws of research methods, which may be both complex and highly variable within different research contexts. Therefore, the Conformity Score metric, which emphasizes technical discussions in review reports, could also serve as a reference for examining whether a review report is AI-generated.

Robustness checks

The review reports in the experimental dataset were already manually checked and rated by MDPI internal editors before we applied the trained model to the dataset. Those editors have qualified education backgrounds for taking the role of an editor and are well-trained for evaluating review report quality according to MDPI review quality assessment guidelines. The editors assigned one of the four rating classes to a review report collected depending on the editors' judgement about a review report's quality. The four rating classes are "Excellent", "Good", "Normal" and "Poor". The predicted review report Conformity Scores could be correlated with those editors' ratings if the Conformity Scores are indicative of a review report's conformity with the MDPI review quality assessment guidelines. The second column of Table 4 shows the average Conformity Scores of each rating class. It can be seen that the average Conformity Scores increase linearly with the ratings enhancing. We infer that the Conformity Score metrics are generally representative of human editors' evaluations of review report quality. To account for the Conformity Score metrics' emphasis on reviewers' "method" discussions, we further examine whether the review reports that are highly/lowly rated by the editors correspond to high/low proportions of "method" discussions. For each rating category, we count the average proportion of the "method" sentences relative to the total number of sentences in a review report, and display the results in Table 4. It is shown in Table 4 that the proportions of "method" discussions increase linearly with the ranks of ratings. Hence the human editors' evaluations'

Table 4 Average Conformity Scores and average proportion of "method" discussions of editors' rating classes. This table reports the average Conformity Scores and the average proportions of "method" discussion sentences of the review reports that are graded by the editors with regard to 4 classes. For a review report, its proportion of "method" discussions is the quotient of the number of sentences classified as "method" divided by the total number of sentences

Rating	Average Conformity Score	Average Proportion of "Method"
Excellent	67.877	53.685%
Good	30.336	51.847%
Normal	14.873	46.969%
Poor	12.048	41.637%

"unconsciously" agree with the proposed principle that "method" discussions should be prioritized on in review evaluating practices.

Since the Conformity Score metrics are an aggregation of sentence count thus it is interesting to explore whether the review reports of high Conformity Scores are simply add-ups of a large number of sentences instead of constructive feedbacks. To address this concern, we sort review reports into 10 groups based on their number of words, ensuring that the review reports in the same group in general have similar numbers of words. For review reports in each group, we show how their proportions of "method" discussions vary with their Conformity Scores in Fig. 14. According to Fig. 14, proportions of "method"

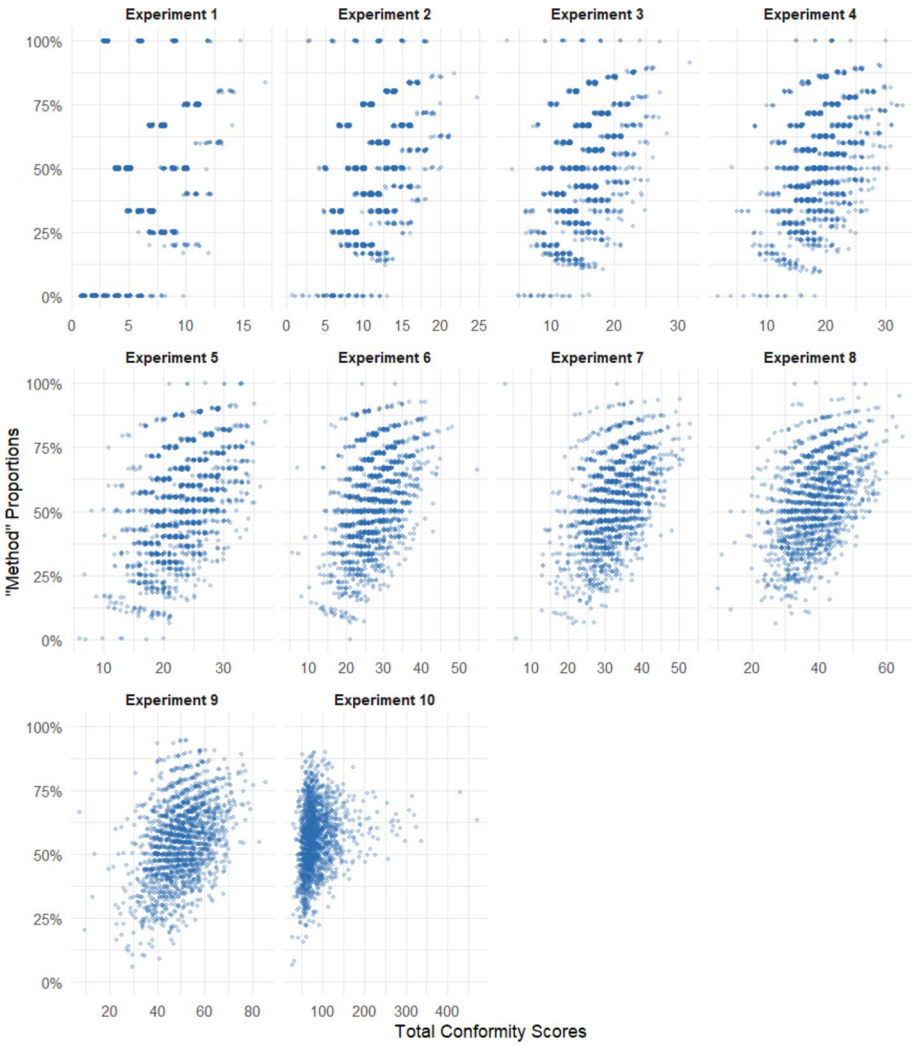


Fig. 14 Conformity scores and Proportions of "Method" discussions of the word count groups. We sort the 18,014 reports in the experimental dataset from according to the word counts of the reports in the ascending order, then we split the dataset into 10 groups of equal size. The bottom/top group includes the reports that have the lowest/highest numbers of word counts

discussions generally have a linear relationship with Conformity Scores. Therefore, in comparison with the review reports of similar length, those review reports of high Conformity Scores indeed have distinguishable proportions of "method" discussions which are considered to be constructive feedbacks.

To conclude, the Conformity Score metrics are generally in line with human evaluations of review report quality and effectively reward those review reports of sufficient technical discussions. However, the way that the Conformity Score metrics is computed allows for manipulation that simply adds up many sentences of superficial comments for a high Conformity Score. Although the counter-argument could be report length still matters such that a long review report full of superficial comments is still valuable than a short review report full of superficial comments, we still advice considering incorporating the other report length adjusted indicators (e.g., proportion of "method" discussions) as a supplementary reference when one benchmarks review report quality using the proposed Conformity Score metrics. For example, for review reports whose proportions of "method" discussions are lower than a certain threshold, one can consider those review reports as low quality review reports.

Conclusion

In this study, we disentangle the textual content of review reports, categorizing the semantic meaning of individual sentences with regard to supervised machine learning models. Based on these models, we classify sentences from review reports with reference to the conformity standards developed by MDPI. This sentence classification lays a foundation for computing the proposed Conformity Score metric, which reflects overall review report quality based on the contribution of each classified sentence in a review report. This metric distinguishes reviewers' technical discussions which, by assumption, contribute significantly to enhancing the quality and impacts of reviewed manuscripts based on other comments. Examining the relationship between the textual features that are commonly regarded as indicators of review report quality and Conformity Scores, we find that the two are generally aligned. We argue that the Conformity Score metrics can serve as a review report assessment benchmark for examining review report contents' conformity with the examination guidelines designed by practitioners.

The novelty of this study is the emphasis on reviewers' technical discussions about reviewed articles. In our view, comprehensive discussions about a study's technical details, such as its methodology, require a deep understanding of the research, and those discussions also directly reflect reviewers' commitment to conducting a thorough review. In addition, some inferences drawn from the examination of the relationship between individual textual features and the Conformity Scores contradict common opinions, shedding light on discussions about indicators of review report quality. It is worth noting that the manual data collecting and labelling process does not depend on the report identifiers as the underlying information used in the modelling, while we still believe that report identifiers should be well-recorded in practice of applying our proposed methodology since report identifiers are critical information enabling report tracing and model validating in iterations of a corpus database.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-026-05539-8.

Acknowledgements The authors would like to thank Dr. Shu-Kun Lin and MDPI for their technical support regarding our study. The authors would also like to thank MDPI Author Services for their professional English Editing.

Author contributions Qiao Sun: Conceptualization, Data curation, Methodology, Software, Supervision, Writing – original draft. Dahai Liu: Data curation, Methodology, Software. Haochen Dong: Conceptualization Methodology, Supervision, Writing – original draft. Shuangfei Li: Validation.

Data availability The review text underlying this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they provide services to MDPI and their cooperation with MDPI did not affect the study design, data collection, and analysis, decision to publish, or the interpretation of the findings. The research was conducted independently and objectively.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Arous, I., Yang, J., Khayati, M., & Cudré-Mauroux, P. (2021). Peer grading the peer reviews: A dual-role approach for lightening the scholarly paper review process. Proceedings of the Web Conference 2021

Ausloos, M., Nedic, O., Fronczak, A., & Fronczak, P. (2016). Quantifying the quality of peer reviewers through Zipf's law. *Scientometrics, 106*, 347–368.

Bharti, P., Ghosal, T., Agarwal, M., & Ekbal, A. (2022). A method for automatically estimating the informativeness of peer reviews. Proceedings of the 19th International Conference on Natural Language Processing (ICON),

Bornmann, L., Wolf, M., & Daniel, H.-D. (2012). Closed versus open reviewing of journal manuscripts: How far do comments differ in language use? *Scientometrics, 91*(3), 843–856.

Bruce, R., Chauvin, A., Trinquart, L., Ravaud, P., & Boutron, I. (2016). Impact of interventions to improve the quality of peer review of biomedical journals: A systematic review and meta-analysis. *BMC Medicine, 14*, 1–16.

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

Callaham, M. L., Baxt, W. G., Waeckerle, J. F., & Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA, 280*(3), 229–231.

Chakraborty, S., Goyal, P., & Mukherjee, A. (2020). Aspect-based sentiment analysis of scientific reviews. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020,

Cheng, L., Bing, L., Yu, Q., Lu, W., & Si, L. (2020). APE: Argument pair extraction from peer review and rebuttal via multi-task learning. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint* arXiv:2004.07180.

D'Arcy, M., Hope, T., Birnbaum, L., & Downey, D. (2024). Marg: Multi-agent review generation for scientific papers. *arXiv preprint* arXiv:2401.04259.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 27*(2), 37–54.

Drori, I., & Te'eni, D. (2024). Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems, 25*(1), 98–109.

Enserink, M. (2001). Peer review and quality: A dubious connection? In: American Association for the Advancement of Science.
Falkenberg, L. J., & Soranno, P. A. (2018). Reviewing reviews: An evaluation of peer reviews of journal article submissions. *Limnology and Oceanography Bulletin, 27*(1), 1–5.
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), Article e0259238.
Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2019). A sentiment augmented deep architecture to predict peer review outcomes. 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL),
Guo, Y., Shang, G., Rennard, V., Vazirgiannis, M., & Clavel, C. (2023). Automatic analysis of substantiation in scientific peer reviews. *arXiv preprint* arXiv:2311.11967.
Henly, S. J., & Dougherty, M. C. (2009). Quality of manuscript reviews in nursing research. *Nursing Outlook, 57*(1), 18–26.
Henly, S. J., Bennett, J. A., & Dougherty, M. C. (2010). Scientific and statistical reviews of manuscripts submitted to Nursing Research: Comparison of completeness, quality, and usefulness. *Nursing Outlook, 58*(4), 188–199.
Hopewell, S., Collins, G. S., Boutron, I., Yu, L. M., Cook, J., Shanyinde, M., Wharton, R., Shamseer, L., & Altman, D. G. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: Retrospective before and after study. *Bmj*. https://doi.org/10.1136/bmj.g4145
Houry, D., Green, S., & Callaham, M. (2012). Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Medical Education, 12*, 1–7.
Hren, D., Pina, D. G., Norman, C. R., & Marušić, A. (2022). What makes or breaks competitive research proposals? A mixed-methods analysis of research grant evaluation reports. *Journal of Informetrics, 16*(2), Article 101289.
Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). Argument mining for understanding peer reviews. *arXiv preprint* arXiv:1903.10104.
Jefferson, T., Rudin, M., Folse, S. B., & Davidoff, F. (2006). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.MR000016.pub2
Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *JAMA, 287*(21), 2786–2790.
Kuznetsov, I., Afzal, O. M., Dercksen, K., Dycke, N., Goldberg, A., Hope, T.,…Leyton-Brown, K. (2024). What Can Natural Language Processing Do for Peer Review? *arXiv preprint* arXiv:2405.06563.
Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., McFarland, D. A., Zou, J., & Yin, Y. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI, 1*(8), Article AIoa2400196.
Luo, J., Feliciani, T., Reinhart, M., Hartstein, J., Das, V., Alabi, O., & Shankar, K. (2021). Analyzing sentiments in peer review reports: Evidence from two science funding agencies. *Quantitative Science Studies, 2*(4), 1271–1295.
Maddi, A., & Miotti, L. (2024). On the peer review reports: Does size matter? *Scientometrics, 129*(10), 5893–5913.
Májovský, M. (2025). AI-generated responses in peer review pose a growing challenge for reviewers and editors: Call for a reviewer rating system. *Available at SSRN 5062076*.
Marrella, D., Jiang, S., Ipaktchi, K., & Liverneaux, P. (2025). Comparing AI-generated and human peer reviews: A study on 11 articles. *Hand Surgery and Rehabilitation*. https://doi.org/10.1016/j.hansur.2025.102225
Meng, J. (2023). Assessing and predicting the quality of peer reviews: A text mining approach. *The Electronic Library, 41*(2/3), 186–203.
Okasa, G., de León, A., Strinzel, M., Jorstad, A., Milzow, K., Egger, M., & Müller, S. (2025). A supervised machine learning approach for assessing grant peer review reports. *Quantitative Science Studies, 6*, 1189–1214.
Park, I.-U., Peacey, M. W., & Munafo, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. *Nature, 506*(7486), 93–96.
Priatna, W. S., Manalu, S. R., & Sundjaja, A. M. (2017). Development of review rating and reporting in open journal system. *Procedia Computer Science, 116*, 645–651.
Rajesh, A., Cloud, G., & Harisinghani, M. G. (2013). Improving the quality of manuscript reviews: Impact of introducing a structured electronic template to submit reviews. *American Journal of Roentgenology, 200*(1), 20–23.
Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18*(2), Article 102946.

Sculley, D., Snoek, J., & Wiltschko, A. (2018). Avoiding a tragedy of the commons in the peer review process. *arXiv preprint* arXiv:1901.06246.

Shattell, M. M., Chinn, P., Thomas, S. P., & Cowling, W. R., III. (2010). Authors' and editors' perspectives on peer review quality in three scholarly nursing journals. *Journal of Nursing Scholarship, 42*(1), 58–65.

Shcherbiak, A., Habibnia, H., Böhm, R., & Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making, 19*, Article e21.

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.

Superchi, C., González, J. A., Solà, I., Cobo, E., Hren, D., & Boutron, I. (2019). Tools used to assess the quality of peer review reports: A methodological systematic review. *BMC Medical Research Methodology, 19*, 1–14.

Superchi, C., Hren, D., Blanco, D., Rius, R., Recchioni, A., Boutron, I., & González, J. A. (2020). Development of ARCADIA: A tool for assessing the quality of peer-review reports in biomedical research. *British Medical Journal Open, 10*(6), Article e035604.

Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity and Peer Review, 5*(1), Article 6.

Thompson, S. R., Agel, J., & Losina, E. (2016). The JBJS peer-review scoring scale: A valid, reliable instrument for measuring the quality of peer review reports. *Learned Publishing, 29*(1), 23–25.

Van Erkel, P. F., & Thijssen, P. (2016). The first one wins: Distilling the primacy effect. *Electoral Studies, 44*, 245–254.

Van Rooyen, S. (2001). The evaluation of peer-review quality. *Learned Publishing, 14*(2), 85–91.

Wang, K., & Wan, X. (2018). Sentiment analysis of peer review texts for scholarly papers. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval,

Wolfram, D., Wang, P., & Abuzahra, F. (2021). An exploration of referees' comments published in open peer review journals: The characteristics of review language and the association between review scrutiny and citations. *Research Evaluation, 30*(3), 314–322.

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? *Journal of Artificial Intelligence Research, 75*, 171–212.

Zhu, L., Lai, Y., Xie, J., Mou, W., Huang, L., Qi, C., Yang, T., Jiang, A., Gan, W., Tang, B., Xiao, M., Chu, G., Liu, Z., Cheng, Q., Lin, A., Luo, P., & Zeng, D. (2025). Evaluating the potential risks of employing large language models in peer review. *Clinical and Translational Discovery, 5*(4), Article e70067.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.