Scientometrics
https://doi.org/10.1007/s11192-026-05603-3

Determining quality dimensions for peer review reports using a Delphi approach

Amanda Sizo[1] · **Adriano Lino**[1,2] · **Álvaro Rocha**[3] · **Luis Paulo Reis**[1]

Received: 1 July 2025 / Accepted: 12 March 2026

Abstract
The quality of peer review reports is essential to the integrity and effectiveness of scholarly communication. Yet review reports are often criticized for being vague, biased, or unconstructive, which limits their usefulness for both authors and editors. Existing frameworks for assessing review quality remain fragmented and are rarely validated through expert consensus. This study aims to define and validate a comprehensive set of quality dimensions for peer review reports, encompassing comments addressed to both authors and editors. We employed a two-phase design combining a thematic analysis of the literature with a Delphi study involving 43 scientific editors, primarily from journals in Computer Science and Engineering. Consensus was reached after two Delphi rounds, resulting in 62 validated statements organized into eight quality dimensions: Helpfulness, Specificity, Fairness, Thoroughness, Courteousness, Readability, Consistency, and Relevance. These findings provide an empirically grounded framework to inform the development of clearer standards for peer review practice.

Keywords Peer-review · Delphi study · Quality of review · Information quality · Quality dimension

Introduction

The quality of a peer review report is recognized as a key indicator of the overall quality of the peer review process (Bruce et al., 2016; Ross-Hellauer et al., 2017). Well-structured review reports help authors improve their research and enable editors to make more informed publishing decisions. However, peer review reports are often criticized for being biased, superficial, or excessively critical (Jamali et al., 2020; Ramachandran et al., 2017). Overly harsh or intentionally vague feedback can make the peer review process particularly

✉ Amanda Sizo
up202010567@edu.fe.up.pt

[1] Artificial Intelligence and Computer Science Laboratory (LIACC), Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias 4200-465, Porto, Portugal

[2] Institute of Engineering and Geosciences, Federal University of Western Pará, Santarém, Brazil

[3] ISEG - Lisbon School of Economics and Management, University of Lisbon, Lisbon, Portugal

challenging for early-career authors, turning what should be a constructive experience into a discouraging one (Wilcox, 2019).

Given this context, journal editors play a central role in identifying and mitigating quality issues within the peer review process. They guide reviewers on the journal's requirements, assess reviewer comments for bias, inconsistencies, or inappropriate content, and provide constructive feedback to improve review quality (Ross-Hellauer et al., 2017). Before forwarding peer review reports to authors, editors evaluate and refine them, addressing any issues as necessary. To support this work, specific tools have been developed to assist editors in evaluating review reports. For example, Sizo et al. (2018) reviewed evaluation methods that include assigning scores to reports based on criteria pre-established by editorial boards, or conducting linguistic analyses of report narratives to assess grammatical, lexical, and semantic features (Sizo et al., 2018). Similarly, a literature review by Superchi et al. (2019) identified 24 tools for assessing review quality and proposed nine quality domains: study relevance, originality, interpretation of results, strengths and weaknesses, manuscript organization, structure and characteristics of reviewer comments, timeliness, and overall usefulness of the review (Superchi et al., 2019). Despite these tools, assessing report quality remains a time-consuming task that often requires manual evaluation. Editors typically rate each report based on its overall quality or on specific criteria they consider relevant.

Poor-quality reports present an information quality (IQ) problem, often due to incompleteness, inconsistency, bias, or lack of clarity, which reduces their usefulness. From a user perspective, information quality problems arise when information fails to meet consumer needs (Ge & Helfert, 2007). Selecting appropriate quality dimensions is fundamental to any information quality-related activity (Batini & Scannapieco, 2016). Efforts have been made to identify quality dimensions for evaluating peer review. Metrics such as reliability, fairness, validity, and efficiency enhance transparency and understanding of the process (Ragone et al., 2013). Thoroughness in peer-review reports has been assessed using text analysis and machine learning, focusing on key content categories such as methods, reporting, results, and study relevance (Severin et al., 2022). Fairness has been studied through language models to detect bias in review texts (Zhang et al., 2022). Constructiveness and politeness have been analyzed using machine learning to classify review sentences and assess tone (Bharti et al., 2024). The helpfulness of peer reviews was initially predicted using SVM regression (Xiong & Litman, 2011), and more recently evaluated with generative AI (Liu et al., 2024).

This study is part of a research project aimed at improving the peer review process, specifically the quality of review reports, and was conducted in phases. The first phase, a systematic literature review, identified quality characteristics and criteria (Sizo et al., 2025). However, merely identifying criteria is insufficient to improve the process from a quality perspective. As Batini and Scannapieco (2016) emphasize, quality depends on the appropriate selection of dimensions. Therefore, the second phase of this project focuses on identifying quality dimensions through a consensus-based method involving experts.

This study aims to develop and refine quality dimensions for peer review reports and validate the statements defining each dimension. The Delphi method was used to achieve consensus among editors on the most relevant dimensions, considering comments directed to both authors and editors. The resulting dimensions serve as a reference for evaluating the quality of peer review reports for original research articles. Consensus among 43 experts was reached in the second round of the Delphi method, leading to the definition of 62 statements across eight quality dimensions. The following sections outline the methodology, techniques, tools, and theoretical framework used to obtain this consensus.

This article is structured as follows: Sect. "Theoretical Foundation" provides the theoretical foundation for the IQ problem in a peer-review context. Sect. "Methods" outlines the methodologies adopted in this study. Sect. "Results" presents the results. Finally, Sect. "Discussion and Conclusion" offers a discussion and conclusion of this study.

Theoretical foundation

Information quality

Information Quality (IQ) is widely recognized as a multidimensional concept and is commonly defined as "fitness for use." (Wang & Strong, 1996). This definition is explicitly user-centered, because the ultimate judge of an information product's quality is its user (Juran, 1992). In practice, IQ is operationalized through measurable dimensions such as accuracy, timeliness, precision, reliability, currency, completeness, and relevance, which provide a basis for monitoring and improving the quality of information products (Wang & Strong, 1996). Some dimensions tend to be broadly applicable across domains (for example, accuracy), while others depend on the application context (for example, efficiency or user satisfaction) (Mattioli et al., 2022). For this reason, IQ research has increasingly adopted a contextual approach, which treats quality as dependent on the situation in which information is produced and used (Shankar & Watts, 2003).

Three main approaches are commonly used to identify and categorize IQ dimensions. First, the theoretical approach derives IQ dimensions from conceptual analyses of information deficiencies and from problems introduced during the "data manufacturing" process. An example is the ontological approach proposed by Wand and Wang, which formalizes inconsistencies between real-world states and their representations in information systems (Wand & Wang, 1996). Second, the intuitive approach derives IQ dimensions from researchers' experience, and the needs of a specific application. Zhu et al. used this approach to assess answer quality in social Q&A, proposing dimensions such as informativeness, politeness, completeness, readability, relevance, conciseness, truthfulness, level of detail, originality, objectivity, novelty, usefulness, and expertise (Zhu et al., 2009). Third, the empirical approach defines IQ dimensions based on their perceived importance to data consumers. Al-Jefri et al. applied this approach in the health domain by eliciting internet users' perceptions to identify health IQ dimensions (Al-Jefri et al., 2018).

Peer review reports can be treated as information products because they convey assessments and recommendations that support editorial decisions and guide author revisions. Under this framing, the quality of peer review reports can be interpreted as IQ in context, where "fitness for use" depends on whether the report provides actionable, decision-relevant, and credible information for its intended users.

Selecting appropriate dimensions is a foundational step in IQ-related work because it determines what is measured and how improvements are evaluated (Batini & Scannapieco, 2016). In the scientific peer-review context, prior work has frequently operationalized report quality using dimensions such as helpfulness, fairness, and thoroughness. Helpfulness concerns whether review comments support revision and improvement, and studies have emphasized the value of automated helpfulness assessment in feedback settings (Patchan et al., 2018). Xiong and Litman (2011) examined whether techniques from product review analysis transfer to peer review for usefulness detection (Xiong & Litman, 2011). Ramachandran et al. (2017) estimated helpfulness from the textual content of reviews using

natural language processing and machine-learning techniques (Ramachandran et al., 2017). Liu et al. (2024) employed generative AI, specifically OpenAI's GPT models, to evaluate review helpfulness through prompt-based text generation (Liu et al., 2024). Fairness refers to objective and impartial evaluation, where author identity attributes should not influence the assessment (Allen et al., 2019). However, reviewers may exhibit bias that affects objective judgment (Phillips, 2011), and empirical work has examined fairness disparities and applied language models to detect bias signals in peer review (Zhang et al., 2022). Thoroughness represents a report that comprehensively assesses a manuscript, covering key components to provide essential insights for both authors and editors (Wang, 2018), and it has been assessed using text analysis and machine learning focused on content categories such as methods, reporting, results, and study relevance (Severin et al., 2022).

This research adopts an empirical approach centered on the editor's perspective to define and delineate the quality dimensions of peer review reports. By involving scientific editors in the validation process, the study aims to capture practical and context-specific criteria that reflect the demands of editorial evaluation.

Peer review and information quality problem

Peer review is well suited for quality improvement, given its crucial role in ensuring the integrity and credibility of scientific research (Golan et al., 2023); however, the process also faces significant challenges that can compromise information quality. Criticisms directed at the peer review process include bias toward certain authors, inability to detect major flaws, unnecessary delays in publication, and inability to uncover corruption/scientific misconduct (Benos et al., 2007). Characterizing the information quality problem in the context of the peer review report involves understanding its purpose, the process's challenges, and stakeholders' perspectives. IQ problems focus on the deficiencies of the information itself. To highlight the main problems, we extracted some criticisms about the content of the peer review report found in the literature. We classified them into user perspective and context-dependent according to the classification of IQ problems (Ge & Helfert, 2007) (see Table 1).

Estimating peer review quality is not straightforward. We must understand the underlying nuances of the peer review texts and the reviewer's intent manifested in those texts (Ghosal et al., 2022). The literature shows some definitions. Superchi (2020) defined "*The quality of a peer- review report is the extent to which a peer- review report helps, first, editors make an informed and unbiased decision about the manuscripts' outcome and, second, authors improve the quality of the submitted manuscript"* (Superchi et al., 2020). Additionally, Ghosal (2022) asserted *"a good peer review should comment on the important sections, address the critical aspects of the paper, perform some definitive roles, while clearly bringing out the reviewer's stand on the work."* (Ghosal et al., 2022). Both definitions are valid for proposing a general conception, but how can all quality attributes expected in a review be achieved objectively?

Although quality is subjective, it is possible to frame it within what the user expects. It is also possible to define specific requirements to achieve a given quality attribute. For example, for a review to be specific, the reviewer should offer precise suggestions, clearly identifying the problem and indicating how it could be addressed, including examples or references when appropriate. When the reviewer fulfills this requirement, the corresponding quality attribute is achieved. Based on this approach, defining a set of statements for each dimension ensures that meeting these statements leads to achieving the desired quality dimension. This reasoning aligns with the definition of quality as "conformance to requirements" (Sahil Sanjeev Salvi,

Table 1 IQ Problems and review problems

IQ problem according to context and user perspective	Evidence of criticisms
The information is not based on fact	"I have seen some evaluative reports written in an angry or rude tone" (Wang, 2018)
The information is of doubtful credibility	"Reviewers do bring their biases into their recommendations due to their social, intellectual, and political consideration" (Wang, 2018)
The information presents an partial view	"the most frequently violated ethical norm is that many referees ask authors to cite the referee's work" (Warner, 2019)
The information is irrelevant to the work	"(…) on occasion reviews are less than helpful" (Provenzale & Stanley, 2006)
The information consists of inconsistent meaning	"The number of manuscript evaluations we have received that convey one opinion of the work to the authors and a different one to the Editors was quite unexpected" (Annesley, 2013) "Some reviewers write a positive assessment of the work in their comments to the authors and then turn around and provide a negative assessment in their private comments to the editors. This is extremely frustrating for editors as it places them in the difficult position of having to reconcile the discrepancy in their reports to the authors." (Rai, 2016)
The information is incomplete	"The Editors are left to wonder whether the omission of comments regarding the basic structure of the work implies that it was well done or that it was never evaluated." (DeMaria, 2003)
The information is compactly represented	"Brevity comments and superficial are not rare. This kind of review is not helpful because is generic and noninformative" (Provenzale & Stanley, 2006) "A very short review suggests to us that little review was actually done; similarly, a 1-sentence condemnation offers little help to the authors" (Cummings & Rivara, 2002)
The information is hard to understand	"My impression has been that these reviewers considered the reviewing job to be a burden and just wanted to get it over. I have found that if there is no statement of an overall reaction from the reviewer, I am sometimes left wondering about what the reviewer really means" (Lee, 1995)

2020). Following this approach, we used the statements identified in the literature review (Sizo et al., 2025) to characterize the expected quality dimensions of a peer review report.

Methods

We conducted a two-step approach: (1) a thematic analysis to develop quality dimensions based on a previous systematic literature review, and (2) a Delphi study to validate these dimensions and assess the importance of the statements with a group of editors from peer-reviewed journals.

Thematic analysis

Based on the results of a previous literature review, eight key quality features were identified for peer review reports: helpful, specific, fair, thorough, readable, courteous, objective, and consistent. The review also uncovered a set of statements related to reviewer comments addressed to authors and editors, as well as statements about the tone and structure of the reports (Sizo et al., 2025). These features were then reconceptualized as quality dimensions, each representing a distinct aspect of review quality. To make these dimensions operational, it was necessary to associate each statement with the dimension it best exemplified, so that the full set of statements would collectively define the expected characteristics of each dimension.

This association process was conducted using a thematic analysis approach (Lucas et al., 2007), which enabled the synthesis of the literature review findings into a structured set of dimensions and statements. The goal was to group statements according to their function, specifically, their role in characterising or supporting the achievement of a given dimension. Two researchers (AS and AL) performed the initial categorization using the dimensions identified in the literature as a starting point. Any disagreements were resolved through consensus or, when needed, with the assistance of additional reviewers (AR or LP). During this process, the Objectivity dimension was found to contain only one relevant statement: "Comments should be focused on the manuscript and research." For conceptual clarity, this was incorporated into the Fairness dimension, which also addresses impartiality (Superchi et al., 2019). Furthermore, a new dimension, Relevance, was introduced to capture statements focused on the significance of the research being reviewed. The final list of dimensions is presented in Table 2.

The Delphi method

The Delphi technique, originally developed by Dalkey and Helmer (1963), is a widely used structured method for eliciting and refining expert judgement through iterative rounds of feedback. It has been applied across diverse fields and for multiple purposes, including exploring complex problems, identifying areas of agreement and disagreement, and supporting both consensus-building and structured dissent (Niederberger et al., 2024).

In many applications, including the present study, Delphi is used as a consensus-oriented survey method, employing successive questionnaire rounds to gather and refine expert input (Giannarou & Zervas, 2014). After each round, participants receive an anonymized summary of the group responses and are invited to reconsider their previous answers in light of this feedback. This iterative process supports reflection and convergence of opinions, although it does not necessarily require complete consensus (Goodman, 1987; Hsu & Sandford, 2007; Keeney et al., 2006). Four key features are generally regarded as central to the Delphi procedure: anonymity, interaction, controlled feedback, and statistical aggregation of group responses (Rowe & Wright, 1999).

Study design

In Round 1, participants independently ranked 62 statements related to the evaluation of peer review report quality across eight dimensions, from both the author and editor perspectives, using a 5-point Likert scale ("essential," "important," "moderately important," "slightly important," "unimportant"). This scale is among the most commonly used

Table 2 Statement by quality dimension

Dimension/Definition	Helpfulness: A review is considered helpful if it supports the editor's decision-making process and assists the authors in improving their manuscript
Statements	1. Provides constructive feedback, with suggestions for manuscript improvement
	2. Provides suggestions for alternative ways to analyze the data
	3. Provides additional comments that add value to the manuscript
	4. Proposes a solution for each problem highlighted
	5. Summarizes the reviewer's interpretation of the study
	6. Identifies major scientific problems or concerns
	7. Clarifies whether concerns are evidence-based or simply hunches
	8. Provides the recommendation regarding publication, whether the paper should be accepted, revised, or rejected
	9. Includes a clear rationale for the recommended decision
	10. Establishes the appropriateness and priority of research for publication
	11. Mentions any omissions
	12. Highlights the amount of work required before the manuscript may be suitable for publication
	13. Indicates whether the topic is important for the journal and whether readers find it interesting
	14. Includes alerts on any ethical concerns (e.g., plagiarism, fraud, unethical research practices)
	15. Identifies areas of the manuscript that the reviewer was unable to adequately assess and suggests other professionals who could be solicited (e.g., statistician)
	16. Suggests a change to a research type that is more appropriate for the content
	17. Describes how the study adds to practice and what it adds to the field
	18. Provides insight to the editor into your approach or engagement with the process
	19. Detects potential conflicts of interest not recognized by the authors
	20. Provides notes if major grammatical errors make the manuscript difficult to read
Dimension	Specificity: A specific recommendation should provide detailed, precise, and actionable feedback. It should clearly outline the necessary improvements and suggest how to implement them, offering concrete proposals for revision
Statements	1. Makes specific suggestions for manuscript improvements
	2. Provides specific comments with a clear explanation of the criticisms and how to resolve them
	3. Provides references and citations from the literature to support the comments
	4. Provides comments detailed enough to assist authors in making necessary modifications
	5. Provides examples to highlight the issues
Dimension	Fairness: A fair review report should adopt a balanced approach, providing constructive critiques that acknowledge both the strengths and weaknesses of the study
Statements	1. Identifies strengths and weaknesses of the manuscript
	2. Identifies both major and minor concerns
	3. Provides general comments, both favourable and negative
	4. Comments should be focused on the manuscript and research
	5. Includes no personal bias in the review
	6. Avoids self-promoting behaviours

Table 2 (continued)

Dimension	Thoroughness: A thorough review addresses all key components of the manuscript in detail. It is typically substantial in length, includes meaningful analyses, and provides a comprehensive evaluation
Statements	1. Provides a review with good length and sufficient details to be useful to both the editor and the author
	2. Provides a detailed analysis of the article sections, as well as tables and figures
	3. Provides notes on whether linguistic editing is needed or not
	4. Evaluates the organization, flow, and readability of the study
	5. Identifies language mistakes and typos
Dimension	Courteousness: The review should be expressed in a polite, respectful, and collegial tone
Statements	1. Is written in a positive tone
	2. Is written in a professional or neutral tone
	3. The tone and vocabulary of the review should be academic
	4. Use respectful and professional language in your written review
	5. Be polite in your suggestions and avoid embarrassing and humiliating comments
Dimension	Readability: The review should be well-organized and clearly structured, ensuring that all information is easy to understand and free from ambiguity. A logical flow allows editors and authors to follow the reviewer's points more effectively
Statements	1. Opening the comments include the title of the manuscript
	2. Comments should start with a summary of the reviewer's interpretation of the work
	3. Provides general comments and specific comments
	4. Provides comments in the order they occur in the manuscript, divided by section
	5. Provides numbered comments
	6. Provides comments organized from the most important to the least important
	7. Provides each new comment in a new paragraph
	8. Is written in a manner that does not reveal the result of the review
	9. Provides comments posed as clear suggestions or observations, instead of in the form of questions
	10. Provides comments organized by theme
	11. Provides comments in the order of the manuscript pages
	12. Reviews should not address the author directly, not as "you". The phrases should be replaced by "the authors" or "the paper"
	13. Provides clear and concise comments
	14. Provides writing referring to page numbers and line numbers in the manuscript
Dimension	Consistency: This dimension assesses the coherence of the review report. Comments addressed to the editor and the author should align with one another and with the reviewer's overall recommendation
Statements	1. Provides consistency between the comments to authors and the recommended decision
	2. Provides comments to the editors congruent with those provided to the authors
Dimension	Relevance: The review should highlight the significance and importance of the study, explaining its contribution to the advancement of knowledge in the field. It should also assess the study's potential impact if published

Table 2 (continued)

Statements	1. Includes statements related to the manuscript's overall contribution to the field
	2. Includes statements related to the potential relevance of the work
	3. Provides an overall assessment of the originality of the study
	4. Include brief notes on the significance of the work and what it adds to current knowledge
	5. Summarizes the overall impression of the manuscript's validity and implications

in Delphi studies (Giannarou & Zervas, 2014). An additional "No opinion" option was included for statements that participants found unclear or believed required further attention. For each statement, participants also had the option to suggest a "Change dimension" if they considered another dimension more appropriate. Each dimension in the survey included a free-text field, allowing participants to elaborate on or explain their responses. Demographic data were also collected in Round 1, including participants' age, gender, highest educational qualification, years of experience as a scientific editor, types of journals they have worked with, and the subject areas of these journals.

In Round 2, participants received an anonymized summary of the full descriptive statistics from the previous round's responses. This allowed them to reconsider and potentially revise their earlier answers based on the collective feedback from all Delphi panel members. Participants also reclassified statements where consensus was not achieved in the previous round. Therefore, each participant received an individualized survey comprising 58 and 57 statements related to comments to the author and editor, respectively, across seven dimensions. Again, for each statement, there was a field for "Change dimension", as well as a free-text field where participants could elaborate on or explain their responses. This process was repeated until a stopping point was reached.

This procedure allowed the dimensional structure itself to be empirically validated throughout the Delphi process. Statements for which participants suggested a different dimension, or which failed to reach the predefined consensus thresholds were systematically reviewed and either reworded, reassigned to a different dimension, or removed between rounds. Only statements that achieved quantitative consensus within their assigned dimension were retained in the final framework. This ensured that the final statement–dimension mapping reflected formal expert agreement rather than being imposed a priori.

Questionnaire development

The structured questionnaire for Round 1 was developed based on the final list of dimensions and statements for writing peer review reports (see Table 2). Questionnaires for subsequent rounds were designed based on the results from the previous round. To meet the study objectives, the survey was divided into eight sections, each corresponding to one of the eight dimensions used to evaluate the quality of peer review reports: Helpfulness, Specificity, Fairness, Thoroughness, Courteousness, Readability, Consistency, and Relevance. Each section included a definition of the respective dimension, enabling participants to assess whether the statements

effectively characterize that specific dimension. Furthermore, statements were categorized by their intended recipient, either the author or the editor, recognizing that these two types of comments serve different roles within the peer review process and therefore require distinct evaluative perspectives.

To support the implementation of the survey rounds, all questionnaires were administered using SurveyJ, a JavaScript library for building web-based surveys and forms. This tool offered a flexible framework that allowed free publication on any web server, along with integrated data storage capabilities. Using this platform, a dedicated web form was developed for participant responses and hosted on GitHub. Prior to deployment, the survey instrument underwent pilot testing with five research fellows experienced in scientific peer review. This stage involved an iterative feedback process aimed at refining the structure and readability of the statements, as well as estimating the average time required to complete the questionnaire.

Survey links were distributed via email, accompanied by reminder messages to encourage participation in each round. To maintain participant anonymity while ensuring response tracking across rounds, each individual was assigned a unique code. Only participants who completed the previous round received access to the next, using the same code for authentication.

Participants' recruitment

A purposive sampling strategy was used to recruit a panel of scientific editors. In this study, editors were regarded as experts in peer review. They oversee the entire review process, managing comments addressed both to the editor and to the authors. They are responsible for making editorial decisions and communicating them to the authors, ensuring that feedback is aligned with the manuscript content and meets authors' expectations.

Delphi studies do not require a fixed number of panelists (Delbecq et al., 1976), although the literature provides guidance on panel size. Hill and Fowles (1975) recommended a minimum of seven participants, while Mitchell and McGoldrick (1994) suggested panel sizes ranging from 10 to 40 participants. Because the Delphi method prioritizes the inclusion of diverse expert perspectives rather than statistical representativeness, a stratified purposive sampling approach was adopted to capture heterogeneity of editorial profiles within the disciplinary scope of the study.

For this study, 474 invitations were sent to editors of journals in the fields of Computer Science and Engineering. The mailing list was compiled through the manual collection of publicly available contact information from journal websites. Therefore, recruitment was deliberately targeted to these disciplinary areas rather than based on a random or cross-disciplinary sample. Given the demanding schedules of editors, a minimum response rate of 10% was anticipated.

Ethical considerations

This study involved expert editors who voluntarily participated in an online Delphi survey. All participants provided informed consent, and no personal or sensitive data were collected beyond basic demographics. The data were anonymized and processed in accordance with data protection regulations (Regulation (EU) 2016/679) (GDPR), the Data Protection Act

2018. At the time of the study, formal ethics approval was not required under our institution's guidelines for low-risk research involving professionals.

Data analyses

Descriptive statistics were used to describe the experts' demographic characteristics and group responses to each statement in each round. To better characterize the panel, we complemented the demographic profile with a journal-quality indicator by assigning each participant the Scimago Journal Rank (SJR) quartile corresponding to the journal in which they serve as an editor.

After each round, results were extracted, organized, and documented in an Excel spreadsheet. Metrics such as mean, median, interquartile range (IQR), standard deviation, and percentage agreement were calculated and reported for every statement.

Consensus for including a statement within a dimension was determined using three combined measures (Giannarou & Zervas, 2014): a) At least 70% of respondents rated the statement between 4 and 5 on a 5-point Likert scale; b) An interquartile range (IQR) of 0 to 2; and c) A standard deviation below 1.5.

To exclude a statement, at least 70% of experts must rate it as 1, accompanied by an IQR of 0 to 2. We excluded statements already agreed upon in the previous round to keep the next round as brief as possible. According to the inclusion and exclusion criteria, we did not include statements in the next round if they achieve consensus for inclusion (> 70% agreement among experts with ratings of 4–5 for each statement) or exclusion (> 70% agreement among experts with ratings of 1 for each statement). This level of agreement has been considered appropriate in previous Delphi studies (Parker et al., 2021). All 'No opinion" responses were excluded from the group response to ensure that each statement's reported percentage agreement or disagreement represented the consensus among only those who felt they knew the answer (Vogel et al., 2019).

All open-ended responses provided by participants across Delphi rounds were systematically reviewed in relation to the corresponding quality dimension. The qualitative analysis focused on identifying suggestions for improving the wording, clarity, and conceptual definition of the dimensions and statements, as well as on detecting recurring concerns or ambiguities raised by participants. These insights were used to refine the questionnaire between rounds. In the Results section, selected excerpts from participants' comments are presented to illustrate the types of feedback received for each dimension. Many responses included personal opinions and professional experiences, which were considered valuable for contextualizing the findings. The complete set of qualitative responses is preserved in the raw dataset.

Results

Participants

A total of 43 participants completed the questionnaire in Round 1 (response rate 11%), and 39 completed it in Round 2 of the Delphi consensus process (response rate 90%). Both rounds had participants of both genders. The median age of the participants was 57 years (interquartile range 13) in Round 1 and 58 years (interquartile range 13) in Round 2. The majority of participants held a PhD-level education (98%). The median number of years of experience working as an editor was 12 (interquartile range 10) in Round 1 and 11

(interquartile range 10) in Round 2. Most participants were editors of specialized, discipline-specific journals, with the most common journal subjects being Computer and Information Sciences and Engineering and Technology. Table 3 presents the demographic characteristics of participants in each round.

Delphi rounds

In Round 1, 62 statements were presented across eight dimensions: Helpfulness (20), Specificity (5), Fairness (6), Thoroughness (5), Courteousness (5), Readability (14),

Table 3 Demographic characteristics of the panel members

Demographics	Round 1 (43) Frequencies	Round 2 (39) Frequencies
Age		
Median	57	58
IQR	49–62 (min–max: 39–82)	49–62 (min–max: 43–82)
Gender		
Female	28% (12)	28% (11)
Male	72% (31)	72% (28)
Degree		
Master's degree	2% (1)	3% (1)
Doctorate degree	98% (41)	97% (37)
Work Experience (Years)		
Median	12	11
IQR	6–16 (min–max: 3–45)	6–16 (min–max: 3–45)
Journal type		
Specialized discipline-specific journal	54% (23)	54% (21)
General multidisciplinary journal	8% (4)	8% (3)
Both	38% (16)	38% (15)
Journal subject		
Computer and Information Sciences	39% (25)	39% (23)
Engineering and Technology	20% (13)	22% (13)
Medical and Health Sciences	14% (9)	14% (8)
Mathematics	8% (5)	8% (5)
Social Sciences	8% (5)	7% (4)
Physical Sciences	5% (3)	3% (2)
Remote sensing	2% (1)	2% (1)
Humanities	2% (1)	2% (1)
Biological Sciences	2% (1)	2% (1)
Multidisciplinary	2% (1)	2% (1)
Journal Quartile		
Q1	62,8% (27)	59% (23)
Q2Q3	37,2% (16)	41% (16)

[‡]The sum of the journal subject represented exceeds the number of participants because participants could select multiple answers

Consistency (2), and Relevance (5). Each statement was evaluated from the perspectives of comments directed to the author and the editor. A total of 31 statements reached consensus regarding comments to the author (29 included and 2 excluded), and 26 reached consensus regarding comments to the editor (23 included and 3 excluded). Statements that achieved consensus were excluded from Round 2. For instance, the "Consistency" dimension reached consensus on all statements from both perspectives and was therefore excluded from Round 2. Twelve statements changed dimensions according to the results of Round 1. These are represented by the statements that moved out of their original dimensions, as shown in Fig. 1. The summary of the results of all statements evaluated in Round 1 is presented in Online Resource.

In the open-ended responses, participants did not suggest modifying the statements. However, they highlighted topics they consider important when writing a review. Most editors showed concern about distinguishing the objectives of comments directed to authors versus those intended for editors, emphasizing the need to avoid repetition. One participant noted, *"The main thing is for the reviewer to be very clear about what the problems may be for both the author and the editor"*. Another added, *"In medical journals, the advice to the author does not need to be repeated in comments to the editor"*. These observations underscore that not all statements are suitable for both audiences. Additionally, one participant remarked that the 'comment to editor' section often seems less significant: *"The section 'comment to editor' should have less weight/importance than the 'comment to author' section because reviewers very often leave it blank."* This perspective can be one reason why most statements had achieved consensus only for comments directed to authors.

Another relevant topic raised concerns the review of language in manuscripts. One participant remarked, *"Language issues should not really be the concern of academic reviewers"*. The study included these statements because they are frequently discussed in publications on manuscript review practices (Jaarsma et al., 2013; Venne, 2015). While it is

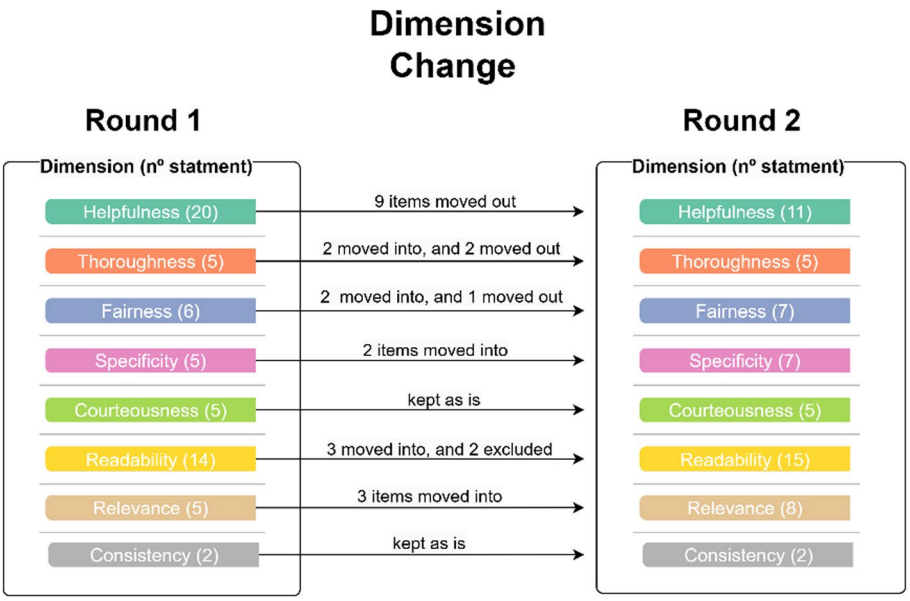

Fig. 1 Number of statements before and after change dimension

understood that such issues are generally less critical than the content and often fall under the editor's responsibility, there are instances where language quality can affect the intelligibility of the work. As noted by Curzon and Cleaton-Jones (2011), when English, spelling, or grammar errors significantly compromise clarity, it is appropriate for the reviewer to comment. Reviewers are encouraged to report poor language, writing style, or grammar when these issues impair the manuscript's overall readability (DiDomenico et al., 2017).

The dimension of courteousness was frequently discussed. One participant noted, *"Sometimes a manuscript can be so dreadful it is hard to be courteous. We should always try, but being honest is also important"*. The literature emphasizes the importance of maintaining a positive, academic, and impartial tone, especially in unfavorable reviews. Regarding tone in rejection, one participant preferred concise and objective explanations stating, *"I would prefer to receive a review that succinctly explains why a manuscript should be rejected but in more courteous terms to the authors. Comments need to be consistent regarding the issues they address but not necessarily in terms of 'tone'."* They also noted that comments to the editor need not align with the tone used for authors or be consistently positive. Another participant warned against an overly positive tone, stating, *"If the feedback sounds okay but the recommendation is to reject, it makes authors feel cheated"*. These perspectives underscore the need for a balance of clarity, honesty, and courtesy in reviewer comments to manage the author's expectations.

In Round 2, we presented 31 statements to comment to the author and 35 to the editor, which had not reached consensus in Round 1. A total of 17 statements reached a consensus regarding comments to the author and 7 regarding comments to the editor. A third round was deemed unnecessary, as all statements reached consensus from at least one perspective, the author or the editor. No statement received a majority vote to change dimensions, and stability of consensus (von der Gracht, 2012) (coefficient of variation $<=0.5$) was achieved between Round 1 and Round 2 for the seven statements that did not reach consensus for comments to either the author or the editor. The summary of the results of all statements evaluated in Round 2 is presented in Online Resource.

From the perspective of comments to authors, consensus was achieved in 100% of the statements related to the dimensions of Fairness, Thoroughness, Courteousness, and Consistency, with the lowest consensus observed in the Relevance dimension. Regarding comments to editors, only the dimensions of Consistency and Relevance reached 100% consensus. For the Courteousness dimension, the statement "*1. Is written in a positive tone.*" did not achieve consensus, aligning with qualitative data indicating that editors do not require a specific tone in their comments, as they prefer to prioritize tone for the authors. The remaining dimensions were less significant, as most of their statements are more relevant to comments to authors. For example, in the Specificity dimension, only one statement "*2. Provides specific comments with a clear explanation of the criticisms and how to resolve them.*" was considered important for editor comments, as the others delve into details related to manuscript critiques. (See Fig. 2).

In summary, 88.7% of the statements reached consensus for at least one type of comment, either author-directed or editor-directed, representing a total of 55 out of 62 statements. From the perspective of comments to authors alone, 48 statements (77.4%) achieved consensus, while 34 statements (54.8%) reached consensus for editor-directed comments (see Fig. 3).

Regarding comments directed to the author, 46 statements were identified as important and representative of the quality of the review report, while three were excluded due to their lack of relevance. For comments to the editor, 22 statements were deemed important, with three excluded for the same reason. A summary of all statements is

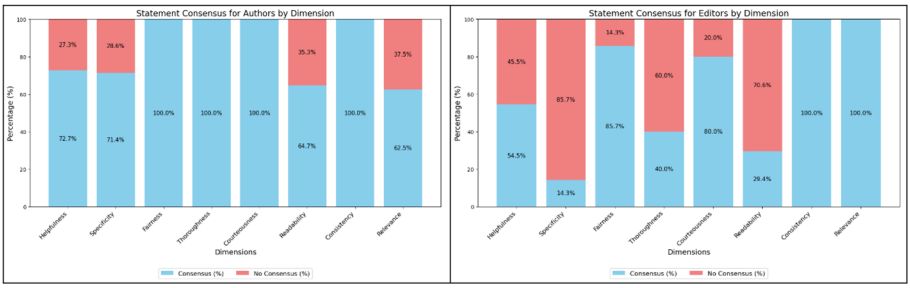

Fig. 2 Quality Dimension Consensus

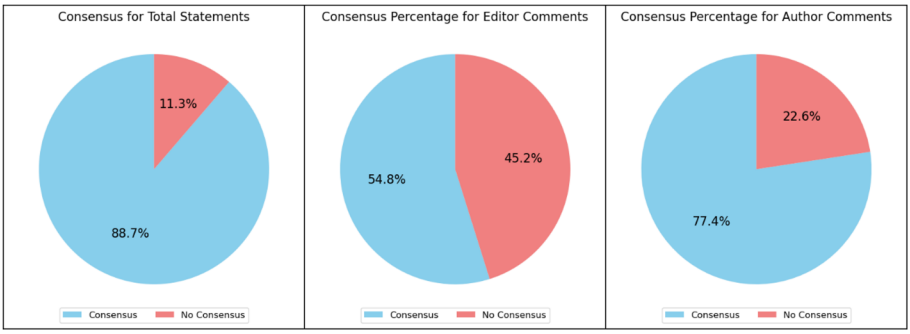

Fig. 3 Consensus by participants

provided Table 4. Detailed descriptive statistics and consensus results are provided in the Appendix.

Additionally, we assessed whether statement ratings varied by journal-quality level by comparing Q1 versus Q2Q3 groups using Mann–Whitney tests, which are appropriate for ordinal Likert data. Across both the "comment to author" and "comment to editor" perspectives, the pattern of results was largely consistent across quartiles, with only a small subset of statements showing nominal differences and limited stability across rounds. This suggests that, in the final converged round, the priorities captured by the Delphi process are broadly similar across journal-quality strata. The test results for all statements evaluated in Rounds 1 and 2 are provided in the Online Resource.

In the open-ended responses, a participant remarked on the Specificity dimension: "*I don't need to have these specific points included in the response to me if they are included in the response to authors.*" Regarding the Thoroughness dimension, another participant stated: *"For example, if a reviewer points out important methodological flaws to the authors, then all I would need in the editor's comments is, 'I recommend this is rejected–see author comments.' But what is really important is that the comments to authors are detailed enough. Sometimes I receive comments as an editor with details about methodological flaws that aren't in the author comments. This makes everyone's job more difficult."* Participants emphasized that reviewers don't need to comment on every aspect of manuscript, focusing instead on points requiring revision. They also noted that if recommendations are detailed enough in the comments to authors, there is no need to repeat

Table 4 Quality dimension and statements consensus

Dim	Comment to author	Comment to editor
Helpfulness	1. Provides constructive feedback, with suggestions for manuscript improvement	1. Provides the recommendation regarding publication, whether the paper should be accepted, revised, or rejected
	2. Summarizes the reviewer's interpretation of the study	2. Highlights the amount of work required before the manuscript may be suitable for publication
	3. Provides additional comments that add value to the manuscript	3. Identifies major scientific problems or concerns
	4. Proposes a solution for each problem highlighted	4. Identifies areas of the manuscript that the reviewer was unable to adequately assess and suggests other professionals who could be solicited (e.g., statistician)
	5. Identifies major scientific problems or concerns	5. Includes a clear rationale for the recommended decision
	6. Includes a clear rationale for the recommended decision	6. Detects potential conflicts of interest not recognized by the authors
	7. Detects potential conflicts of interest not recognized by the authors	
Specificity	1. Makes specific suggestions for manuscript improvements	1. Provides specific comments with a clear explanation of the criticisms and how to resolve them
	2. Provides specific comments with a clear explanation of the criticisms and how to resolve them	
	3. Provides references and citations from the literature to support the comments	
	4. Provides comments detailed enough to assist authors in making necessary modifications	
	5. Provides examples to highlight the issues	

Table 4 (continued)

Dim	Comment to author	Comment to editor
Fairness	1. Identifies strengths and weaknesses of the manuscript	1. Identifies strengths and weaknesses of the manuscript
	2. Clarifies whether concerns are evidence-based or simply hunches	2. Clarifies whether concerns are evidence-based or simply hunches
	3. Provides general comments, both favorable and negative	3. Comments should be focused on the manuscript and research
	4. Comments should be focused on the manuscript and research	4. Includes no personal bias in the review
	5. Includes no personal bias in the review	5. Avoids self-promoting behaviors
	6. Avoids self-promoting behaviors	6. Includes alerts on any ethical concerns (e.g., plagiarism, fraud, unethical research practices)
	7. Includes alerts on any ethical concerns (e.g., plagiarism, fraud, unethical research practices)	
Thoroughness	1. Provides a review with good length and sufficient details to be useful to both the editor and the author	1. Identifies both major and minor concerns
	2. Identifies both major and minor concerns	2. Mentions any omissions
	3. Provides a detailed analysis of the article sections, as well as tables and figures	
	4. Evaluates the organization, flow, and readability of the study	
	5. Mentions any omissions	
Courteousness	1. Is written in a positive tone	1. Is written in a professional or neutral tone
	2. Is written in a professional or neutral tone	2. The tone and vocabulary of the review should be academic
	3. The tone and vocabulary of the review should be academic	3. Use respectful and professional language in your written review
	4. Use respectful and professional language in your written review	4. Be polite in your suggestions and avoid embarrassing and humiliating comments
	5. Be polite in your suggestions and avoid embarrassing and humiliating comments	

Table 4 (continued)

Dim	Comment to author	Comment to editor
Readability	1. Comments should start with a summary of the reviewer's interpretation of the work	1. Provides notes if major grammatical errors make the manuscript difficult to read
	2. Provides general comments and specific comments	2. Provides clear and concise comments
	3. Provides comments in the order they occur in the manuscript, divided by section	
	4. Provides numbered comments	
	5. Provides comments organized from the most important to the least important	
	6. Provides each new comment in a new paragraph	
	7. Provides notes if major grammatical errors make the manuscript difficult to read	
	8. Provides clear and concise comments	
	9. Provides writing referring to page numbers and line numbers in the manuscript	
Consistency	1. Provides consistency between the comments to authors and the recommended decision	1. Provides consistency between the comments to authors and the recommended decision
	2. Provides comments to the editors congruent with those provided to the authors	2. Provides comments to the editors congruent with those provided to the authors

Table 4 (continued)

Dim	Comment to author	Comment to editor
Relevance	1. Includes statements related to the manuscript's overall contribution to the field	1. Includes statements related to the manuscript's overall contribution to the field
	2. Provides an overall assessment of the originality of the study	2. Includes statements related to the potential relevance of the work
	3. Include brief notes on the significance of the work and what it adds to current knowledge	3. Provides an overall assessment of the originality of the study
	4. Summarizes the overall impression of the manuscript's validity and implications	4. Include brief notes on the significance of the work and what it adds to current knowledge
	5. Describes how the study adds to practice and what it adds to the field	5. Summarizes the overall impression of the manuscript's validity and implications
		6. Establishes the appropriateness and priority of research for publication
		7. Indicates whether the topic is important for the journal and whether readers find it interesting
		8. Describes how the study adds to practice and what it adds to the field

them for the editor. On courteousness, one participant wrote: *"As I think about it, a review doesn't need to be written in a 'positive' tone to be courteous. Perhaps 'positive' is not the best word. 'Non-insulting' is more key."* It was noted that a "positive tone" can be interpreted differently across individuals and cultures and may not always be appropriate, such as in studies involving methods perceived as cruel to animals. Indeed, the statement "*1. Is written in a positive tone.*" proved controversial and was the only phrase in this dimension that failed to reach consensus for comments to editors. On readability, editors noted: "*Some of these suggestions are incompatible: you can only organize the comments in one order, so by page number means not by importance.*" While the statements may be mutually exclusive, they define the dimension, leaving it up to the editor to decide how they prefer the review to be organized. Additionally, another participant stated: "*We don't ask for numbered comments to improve the readability of the review. We ask for these so authors can refer to specific reviewer comments in the rebuttal letter."* This highlights that the organization of the review is more useful for comments directed to authors.

Discussion and conclusion

By integrating evidence from literature with expert consensus, the framework contributes to a more transparent and actionable understanding of what constitutes a high-quality peer review report. This study introduces eight quality dimensions, Helpfulness, Specificity, Fairness, Thoroughness, Courteousness, Readability, Consistency, and Relevance, for assessing the quality of peer review reports. To our knowledge, this is the first Delphi study to establish expert consensus among scientific editors regarding which dimensions best characterize review quality. Through a two-round Delphi process, we validated 62 statements that operationalize these dimensions and clarified their applicability to comments addressed to authors and to editors.

The study addresses a well-documented gap in the literature: while peer review quality is frequently discussed in normative terms, few studies provide empirically grounded and operationalizable criteria. Our findings offer a structured and concrete framework that translates abstract principles such as fairness, thoroughness, and helpfulness into specific, observable characteristics of review reports. This level of granularity supports both reviewers, by clarifying expectations, and editors, by offering more consistent criteria for evaluating review quality.

For instance, to ensure thoroughness, a reviewer should assess key components of a manuscript, including the materials and methods, presentation and reporting, results and discussion, and the overall significance of the study (Severin et al., 2022). Our statements further specify that a thorough review should be sufficiently detailed to address both major sections and minor points, as well as the overall readability of the manuscript. Likewise, fairness involves more than the absence of bias. According to our framework, reviewers are expected to differentiate between evidence-based concerns and subjective opinions, acknowledge both strengths and weaknesses, and avoid self-promotion. By introducing this level of granularity, the framework helps prevent common pitfalls in peer review (Zhang et al., 2022) translating abstract principles into concrete and actionable reviewer behaviors. In addition, features such as "mentions a problem," "offers a suggestion," "provides examples," and "includes praise" have been associated with the perceived helpfulness of review comments (Liu et al., 2024). However, our findings indicate that "offering suggestions" and "providing examples" are not merely general markers of helpfulness, they

are key indicators of the Specificity dimension. These elements are particularly important for ensuring that reviewers deliver actionable recommendations to address a manuscript's weaknesses, thereby increasing the practical value of their feedback.

The results also allow us to identify which quality dimensions are perceived as most important depending on the target of the review comments. For comments addressed to editors, all statements related to the dimensions of Consistency and Relevance reached consensus. This finding aligns with existing evidence indicating that editors value coherence between comments to authors, confidential comments to editors, and the final recommendation. When discrepancies occur, editors often need to intervene to clarify or correct the feedback before communicating a decision to authors, in order to avoid misunderstandings. Relevance, which includes the ability to highlight the significance and contribution of the study, is likewise recognized as a key criterion in editorial decision-making and manuscript acceptance.

For comments addressed to authors, the dimensions of Fairness, Courteousness, and Thoroughness, in addition to Consistency, achieved 100% consensus across all statements. These results are consistent with prior research showing that editors value reviews that are balanced, respectful in tone, and sufficiently detailed to support authors in improving their manuscripts. Together, these findings reinforce the validity of the proposed framework and its alignment with established expectations in scholarly peer review.

Future work

Future research should explore how the validated quality dimensions and statements proposed in this study can be operationalized in computational systems for the automated or semi-automated assessment of peer review reports. The framework offers structured criteria that may serve as inputs for applications based on Natural Language Processing (NLP) and Large Language Models (LLMs). For example, future studies could investigate the use of sentiment analysis and tone detection techniques to operationalize the Courteousness dimension, building on prior work that has examined praise, criticism, politeness, and constructiveness in review texts (Bharti et al., 2023; Thelwall, 2020).

Similarly, future work could assess whether information extraction techniques, such as Named Entity Recognition and transformer-based models (e.g., BERT), are effective in capturing elements related to the Specificity dimension, including references to manuscript sections, figures, or tables. In addition, consistency between comments directed to authors and those directed to editors could be explored using text similarity approaches, such as Sentence-BERT (Nils Reimers, 2019), to identify potential discrepancies in reviewer feedback.

Beyond computational applications, future studies may investigate the integration of the validated statements into editorial guidelines, reviewer training materials, automated screening tools, and prompt design for LLM-based support systems. Such developments could contribute to improving the efficiency, transparency, and overall quality of the peer review process for both editors and authors. These applications were not empirically tested in the present study and therefore remain directions for future investigation. Empirical validation of AI-based implementations, including their reliability and practical usefulness in editorial contexts, represents an important next step.

Limitations

There are several limitations to this study. First, the number of editors who completed all Delphi rounds was relatively small, which may limit the generalizability of the findings, particularly with respect to editorial perspectives. This constraint is common in Delphi research, where sample sizes are typically limited and results are not intended to be statistically representative of a broader population (Fink-Hafner et al., 2019). Nevertheless, the Delphi method is well suited to addressing complex issues that are difficult to capture using conventional empirical approaches (Vernon & Vernon, 2013), and the findings provide a useful foundation for the development and refinement of peer review guidelines.

Second, the recruitment strategy relied on a purposive sample of editors from journals in the fields of Computer Science and Engineering, which led to the underrepresentation of other disciplines. Peer review practices vary considerably across fields and publication venues (Horbach & Halffman, 2018), and the perceived importance of certain quality criteria may therefore differ by disciplinary context. As a result, the findings primarily reflect editorial expectations within STEM domains and require further validation in the contexts of social sciences, humanities, and medical publishing.

Appendix

See (Table 5).

Table 5 Descriptive statistics and consensus results

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
Helpfulness	1. Provides constructive feedback, with suggestions for manuscript improvement	Round 1 Included	100% (4/5)	0	5	1	0,3937	0,1550	5	5	No Consensus	38% (4/5)	2	3	4	1,1424	1,3050	3	3	Help: 92,31% Spec: 2,56% Fair: 0% Thor: 5,13% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	2. Summarizes the reviewer's interpretation of the study	Round 2 Included	74% (4/5)	2	5	4	1,0990	1,2078	4	4	No Consensus	23% (4/5)	1	3	4	1,1094	1,2308	3	3	Help: 94,87% Spec: 2,56% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	3. Provides additional comments that add value to the manuscript	Round 1 Included	91% (4/5)	1	4	2	0,6444	0,4153	4	4	No Consensus	32% (4/5)	2	3	4	1,2625	1,5939	3	3	Help: 82,05% Spec: 5,13% Fair: 0% Thor: 10,26% Cour: 0% Cons: 0% Rele: 2,56% Read: 0%
	4. Proposes a solution for each problem highlighted	Round 2 Included	85% (4/5)	1	5	4	1,0087	1,0175	5	4	No Consensus	23% (4/5)	1	3	4	1,2635	1,5965	3	3	Help: 79,49% Spec: 10,26% Fair: 0% Thor: 10,26% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	5. Provides the recommendation regarding publication, whether the paper should be accepted, revised, or rejected	No Consensus	18% (4/5)	2	1	4	1,3829	1,9125	2	2	Round 1 Included	93% (4/5)	0	5	4	0,9765	0,9535	5	5	Help: 92,31% Spec: 2,56% Fair: 0% Thor: 0% Cour: 5,13% Cons: 0% Rele: 0% Read: 0%
	6. Highlights the amount of work required before the manuscript may be suitable for publication	No Consensus	39% (4/5)	2	4	4	1,3950	1,9459	3	3	Round 2 Included	74% (4/5)	2	5	4	1,1459	1,3131	4	4	Help: 92,31% Spec: 2,56% Fair: 0% Thor: 2,56% Cour: 0% Cons: 2,56% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	7. Identifies major scientific problems or concerns	Round 1 Included	100% (4/5)	0	5	1	0,4275	0,1827	5	5	Round 1 Included	84% (4/5)	1	5	4	1,1864	1,4075	5	4	Help: 37,21% Spec: 30,23% Fair: 0% Thor: 23,26% Cour: 0% Cons: 4,65% Rele: 4,65% Read: 0%
	8. Identifies areas of the manuscript that the reviewer was unable to adequately assess and suggests other professionals who could be solicited (e.g., statistician)	No Consensus	15% (4/5)	2	1	4	1,3367	1,7868	2	2	Round 1 Included	72% (4/5)	2	5	3	1,0116	1,0233	4	4	Help: 94,87% Spec: 0% Fair: 2,56% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	9. Provides insight to the editor into your approach or engagement with the process	Round 1 Excluded	70% (1)	2	1	4	1,3845	1,9169	1	2	No Consensus	51% (4/5)	2	3	4	1,1406	1,3009	4	4	Help: 100% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	10. Includes a clear rationale for the recommended decision	Round 1 Included	72% (4/5)	2	5	4	1,5049	2,2647	4	4	Round 1 Included	95% (4/5)	1	5	4	0,9077	0,8239	5	5	Help: 41,86% Spec: 13,95% Fair: 11,63% Thor: 11,63% Cour: 0% Cons: 4,65% Rele: 16,28% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	11. Detects potential conflicts of interest not recognized by the authors	Round 1 Included	78% (4/5)	1	5	4	1,0854	1,1780	4	4	Round 1 Included	81% (4/5)	1	5	4	1,1859	1,4064	5	4	Help: 65,12% Spec: 13,95% Fair: 4,65% Thor: 9,3% Cour: 0% Cons: 2,33% Rele: 0% Read: 4,65%
Specificity	1. Makes specific suggestions for manuscript improvements	Round 1 Included	93% (4/5)	1	5	2	0,6127	0,3754	5	5	No Consensus	33% (4/5)	2	2	4	1,4046	1,9730	3	3	Help: 0% Spec: 97,44% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	2. Provides specific comments with a clear explanation of the criticisms and how to resolve them	Round 1 Included	90% (4/5)	1	5	2	0,6713	0,4506	5	5	Round 1 Included	74% (4/5)	2	5	4	1,4299	2,0447	4	4	Help: 0% Spec: 90,7% Fair: 0% Thor: 4,65% Cour: 0% Cons: 0% Rele: 4,65% Read: 0%
	3. Provides references and citations from the literature to support the comments	Round 2 Included	92% (4/5)	1	5	3	0,7180	0,5155	5	5	No Consensus	26% (4/5)	3	1	4	1,4486	2,0985	2	3	Help: 0% Spec: 97,44% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	4. Provides comments detailed enough to assist authors in making necessary modifications	Round 1 Included	84% (4/5)	1	5	2	0,7668	0,5880	5	4	No Consensus	26% (4/5)	3	1	4	1,4486	2,0985	2	2	Help: 0% Spec: 97,44% Fair: 0% Thor: 0% Cour: 2,56% Cons: 0% Rele: 0% Read: 0%
	5. Provides examples to highlight the issues	Round 2 Included	82% (4/5)	1	5	4	1,1118	1,2362	5	4	No Consensus	21% (4/5)	2	1	4	1,3533	1,8313	2	2	Help: 0% Spec: 100% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	6. Suggests a change to a research type that is more appropriate for the content	No Consensus	37% (4/5)	3	5	4	1,4242	2,0284	3	3	No Consensus	31% (4/5)	2	2	4	1,2967	1,6815	3	3	Help: 0% Spec: 100% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	7. Provides suggestions for alternative ways to analyze the data	No Consensus	41% (4/5)	2	3	4	1,2245	1,4993	3	3	No Consensus	28% (4/5)	2	3	4	1,3215	1,7463	3	3	Help: 0% Spec: 97,44% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 0%
Fairness	1. Identifies strengths and weaknesses of the manuscript	Round 1 Included	74% (4/5)	2	5	4	1,2150	1,4762	4	4	Round 1 Included	70% (4/5)	2	5	4	1,3244	1,7542	4	4	Help: 9,3% Spec: 0% Fair: 67,44% Thor: 16,28% Cour: 0% Cons: 4,65% Rele: 2,33% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	2. Clarifies whether concerns are evidence-based or simply hunches	Round 1 Included	86% (4/5)	1	4	3	0,8117	0,6589	4	4	Round 1 Included	70% (4/5)	2	5	4	1,1850	1,4042	4	4	Help: 39,53% Spec: 6,98% Fair: 44,19% Thor: 4,65% Cour: 2,33% Cons: 0% Rele: 2,33% Read: 0%
	3. Provides general comments, both favorable and negative	Round 2 Included	82% (4/5)	1	5	4	0,9775	0,9555	5	4	No Consensus	51% (4/5)	2	3	4	1,2425	1,5439	4	4	Help: 0% Spec: 0% Fair: 94,87% Thor: 5,13% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	4. Comments should be focused on the manuscript and research	Round 1 Included	93% (4/5)	0	5	2	0,5990	0,3588	5	5	Round 1 Included	72% (4/5)	2	5	4	1,4879	2,2137	5	4	Help: 6,98% Spec: 0% Fair: 76,74% Thor: 0% Cour: 4,65% Cons: 0% Rele: 11,63% Read: 0%
	5. Includes no personal bias in the review	Round 1 Included	100% (4/5)	1	5	1	0,4450	0,1980	5	5	Round 1 Included	86% (4/5)	1	5	4	1,0833	1,1736	5	4	Help: 0% Spec: 0% Fair: 100% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	6. Avoids self-promoting behaviors	Round 1 Included	81% (4/5)	1	5	4	0,9423	0,8879	5	4	Round 1 Included	74% (4/5)	2	5	4	1,2410	1,5401	5	4	Help: 0% Spec: 0% Fair: 90,7% Thor: 0% Cour: 4,65% Cons: 0% Rele: 4,65% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	7. Includes alerts on any ethical concerns (e.g., plagiarism, fraud, unethical research practices)	Round 1 Included	86% (4/5)	0	5	4	0,8652	0,7486	5	5	Round 1 Included	100% (4/5)	0	5	1	0,3244	0,1052	5	5	Help: 37,21% Spec: 9,3% Fair: 39,53% Thor: 6,98% Cour: 0% Cons: 6,98% Rele: 0% Read: 0%
Thoroughness	1. Provides a review with good length and sufficient details to be useful to both the editor and the author	Round 1 Included	88% (4/5)	1	5	3	0,8830	0,7796	5	4	No Consensus	44% (4/5)	2	3	4	1,2293	1,5111	3	4	Help: 2,56% Spec: 0% Fair: 0% Thor: 97,44% Cour: 0% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	2. Identifies both major and minor concerns	Round 1 Included	86% (4/5)	1	5	3	0,7875	0,6202	5	4	Round 2 Included	84% (4/5)	1	5	4	1,0882	1,1842	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 100% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	3. Provides a detailed analysis of the article sections, as well as tables and figures	Round 2 Included	87% (4/5)	0	5	4	1,1209	1,2564	5	5	No Consensus	23% (4/5)	2	1	4	1,4289	2,0418	2	2	Help: 0% Spec: 0% Fair: 0% Thor: 100% Cour: 0% Cons: 0% Rele: 0% Read: 0%
	4. Evaluates the organization, flow, and readability of the study	Round 2 Included	92% (4/5)	0	5	2	0,6047	0,3657	5	5	No Consensus	31% (4/5)	2	2	4	1,3799	1,9042	2	3	Help: 0% Spec: 0% Fair: 0% Thor: 97,44% Cour: 0% Cons: 0% Rele: 0% Read: 2,56%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	5. Mentions any omissions	Round 1 Included	74% (4/5)	2	5	3	0,9053	0,8195	4	4	Round 1 Included	74% (4/5)	1	4	4	1,3253	1,7564	4	4	Help: 34,88% Spec: 0% Fair: 0% Thor: 44,19% Cour: 0% Cons: 6,98% Rele: 13,95% Read: 0%
Courteousness	1. Is written in a positive tone	Round 2 Included	74% (4/5)	2	5	4	1,0631	1,1302	5	4	No Consensus	51% (4/5)	2	5	4	1,2022	1,4453	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 100% Cons: 0% Rele: 0% Read: 0%
	2. Is written in a professional or neutral tone	Round 1 Included	91% (4/5)	1	5	2	0,6597	0,4352	5	5	Round 1 Included	72% (4/5)	2	5	4	1,3452	1,8095	5	4	Help: 0% Spec: 0% Fair: 4,65% Thor: 0% Cour: 95,35% Cons: 0% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	3. The tone and vocabulary of the review should be academic	Round 1 Included	77% (4/5)	1	5	2	0,7945	0,6312	4	4	Round 1 Included	72% (4/5)	2	5	4	1,1628	1,3522	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 97,67% Cons: 0% Rele: 2,33% Read: 0%
	4. Use respectful and professional language in your written review	Round 1 Included	91% (4/5)	1	5	2	0,6656	0,4430	5	5	Round 1 Included	77% (4/5)	1	5	4	1,3837	1,9147	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 97,67% Cons: 0% Rele: 2,33% Read: 0%
	5. Be polite in your suggestions and avoid embarrassing and humiliating comments	Round 1 Included	88% (4/5)	0	5	3	0,7636	0,5830	5	5	Round 1 Included	74% (4/5)	2	5	4	1,3615	1,8537	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 97,67% Cons: 0% Rele: 2,33% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
Read-ability	1. Opening the comments include the title of the manuscript	Round 1 Excluded	71% (1)	1	1	4	1,2037	1,4488	1	2	Round 1 Excluded	71% (1)	1	1	4	1,2601	1,5878	1	2	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%
	2. Provides notes on whether linguistic editing is needed or not	No Consensus	67% (4/5)	2	5	4	1,3676	1,8704	4	4	No Consensus	64% (4/5)	2	5	4	1,2005	1,4413	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 5,13% Cour: 0% Cons: 0% Rele: 0% Read: 94,87%
	3. Comments should start with a summary of the reviewer's interpretation of the work	Round 2 Included	82% (4/5)	1	5	4	1,1773	1,3860	5	4	No Consensus	32% (4/5)	2	3	4	1,4175	2,0092	3	3	Help: 0% Spec: 0% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 97,44%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	4. Provides general comments and specific comments	Round 2 Included	90% (4/5)	0	5	3	0,8148	0,6640	5	5	No Consensus	41% (4/5)	3	3	4	1,3322	1,7746	3	3	Help: 0% Spec: 2,56% Fair: 0% Thor: 2,56% Cour: 0% Cons: 0% Rele: 0% Read: 94,87%
	5. Provides comments in the order they occur in the manuscript, divided by section	Round 2 Included	85% (4/5)	1	5	4	1,1406	1,3009	5	4	No Consensus	45% (4/5)	2	5	4	1,4736	2,1714	3	3	Help: 2,56% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 97,44%
	6. Provides numbered comments	Round 2 Included	82% (4/5)	1	5	4	1,2452	1,5506	5	4	No Consensus	31% (4/5)	2	3	4	1,2667	1,6046	3	3	Help: 2,56% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 97,44%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	7. Provides comments organized from the most important to the least important	Round 2 Included	74% (4/5)	2	5	4	1,4954	2,2362	5	4	No Consensus	46% (4/5)	2	3	4	1,4133	1,9973	3	3	Help: 2,56% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 97,44%
	8. Provides each new comment in a new paragraph	Round 2 Included	76% (4/5)	1	5	4	1,3788	1,9011	5	4	No Consensus	26% (4/5)	2	3	4	1,2520	1,5676	3	3	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%
	9. Identifies language mistakes and typos	No Consensus	49% (4/5)	2	5	4	1,3498	1,8219	3	4	No Consensus	28% (4/5)	2	3	4	1,3533	1,8313	3	3	Help: 2,56% Spec: 0% Fair: 0% Thor: 5,13% Cour: 0% Cons: 0% Rele: 0% Read: 92,31%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	10. Is written in a manner that does not reveal the result of the review	Round 1 Excluded	70% (1)	1	1	4	1,5053	2,2658	1	2	Round 1 Excluded	86% (1)	0	1	4	1,1798	1,3920	1	1	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 2,33% Cons: 0% Rele: 0% Read: 97,67%
	11. Provides comments posed as clear suggestions or observations, instead of in the form of questions	No Consensus	38% (4/5)	3	3	4	1,3443	1,8070	3	3	No Consensus	28% (4/5)	3	1	4	1,4639	2,1430	3	3	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%
	12. Provides comments organized by theme	No Consensus	28% (4/5)	2	3	4	1,2967	1,6815	3	3	No Consensus	21% (4/5)	2	3	4	1,2731	1,6208	3	3	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	13. Provides comments in the order of the manuscript pages	No Consensus	68% (4/5)	2	5	4	1,2901	1,6643	4	4	No Consensus	39% (4/5)	2	3	4	1,4051	1,9744	3	3	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%
	14. Reviews should not address the author directly, not as "you". The phrases should be replaced by "the authors" or "the paper"	No Consensus	29% (4/5)	2	2	4	1,3726	1,8841	3	3	Round 1 Excluded	71% (1)	1	1	4	1,1153	1,2439	1	2	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 2,56% Cons: 0% Rele: 0% Read: 97,44%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	15. Provides notes if major grammatical errors make the manuscript difficult to read	Round 1 Included	70% (4/5)	2	5	3	0,9956	0,9911	4	4	Round 2 Included	77% (4/5)	1	5	4	0,9423	0,8880	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 5,13% Cour: 0% Cons: 0% Rele: 0% Read: 94,87%
	16. Provides clear and concise comments	Round 1 Included	74% (4/5)	2	5	4	1,3274	1,7619	5	4	Round 1 Included	72% (4/5)	2	5	4	1,2798	1,6379	5	4	Help: 4,65% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 2,33% Rele: 2,33% Read: 90,7%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	17. Provides writing referring to page numbers and line numbers in the manuscript	Round 2 Included	82% (4/5)	1	5	4	1,0416	1,0850	5	4	No Consensus	56% (4/5)	2	5	4	1,4046	1,9730	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 0% Read: 100%
Consistency	1. Provides consistency between the comments to authors and the recommended decision	Round 1 Included	86% (4/5)	0	5	4	1,3263	1,7590	5	4	Round 1 Included	88% (4/5)	1	5	4	1,2506	1,5639	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 100% Rele: 0% Read: 0%
	2. Provides comments to the editors congruent with those provided to the authors	Round 1 Included	86% (4/5)	0	5	4	1,2209	1,4906	5	4	Round 1 Included	83% (4/5)	1	5	4	1,0373	1,0761	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 100% Rele: 0% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
Relevance	1. Includes statements related to the manuscript's overall contribution to the field	Round 1 Included	72% (4/5)	1	4	4	1,0364	1,0742	4	4	Round 1 Included	72% (4/5)	2	5	4	1,2711	1,6157	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%
	2. Includes statements related to the potential relevance of the work	No Consensus	67% (4/5)	2	5	4	1,0634	1,1309	4	4	Round 2 Included	90% (4/5)	1	5	4	0,8545	0,7301	5	5	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%
	3. Provides an overall assessment of the originality of the study	Round 1 Included	86% (4/5)	1	4	4	0,9053	0,8195	4	4	Round 1 Included	86% (4/5)	1	4	4	1,0370	1,0753	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	4. Include brief notes on the significance of the work and what it adds to current knowledge	Round 1 Included	74% (4/5)	1	4	4	1,1305	1,2780	4	4	Round 1 Included	77% (4/5)	1	4	4	1,3140	1,7265	4	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%
	5. Summarizes the overall impression of the manuscript's validity and implications	Round 2 Included	72% (4/5)	2	5	4	1,1921	1,4211	4	4	Round 2 Included	77% (4/5)	1	5	4	1,4770	2,1816	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%
	6. Establishes the appropriateness and priority of research for publication	No Consensus	31% (4/5)	3	2	4	1,5681	2,4588	2	3	Round 2 Included	92% (4/5)	1	5	3	0,7929	0,6287	5	5	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%

Table 5 (continued)

Dim	Statement	Comment to Author									Comment to Author									Dimension Agree
		Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	Consensu	Agree	IQR	Mo	Range	Std	Var	Md	Mean	
	7. Indicates whether the topic is important for the journal and whether readers find it interesting	No Consensus	33% (4/5)	2	1	4	1,4831	2,1997	3	3	Round 2 Included	87% (4/5)	1	5	4	1,1203	1,2551	5	4	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%
	8. Describes how the study adds to practice and what it adds to the field	Round 2 Included	72% (4/5)	2	5	4	1,2223	1,4939	4	4	Round 2 Included	95% (4/5)	1	5	3	0,7475	0,5587	5	5	Help: 0% Spec: 0% Fair: 0% Thor: 0% Cour: 0% Cons: 0% Rele: 100% Read: 0%

IQR: 0 = high consensus; 1 = good consensus; 2 = poor consensus

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-026-05603-3.

Acknowledgements We extend our thanks to all the editors who participated in this Delphi study and contributed to the development of guidelines for writing peer review reports.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by FCT–Fundação para a Ciência e Tecnologia, I.P. by project reference and DOI identifierhttps://doi.org/10.54499/2022.13474.BD. This work was financially supported by: UID/00027/2025 of the LIACC–Artificial Intelligence and Computer Science Laboratory withhttps://doi.org/10.54499/UID/00027/2025, funded by Fundação para a Ciência e a Tecnologia, I.P./ MECI through the national funds.

Declarations

Conflict of interest The author(s) declare no competing interests.

References

Al-Jefri, M., Evans, R., Uchyigit, G., & Ghezzi, P. (2018). What is health information quality? Ethical dimension and perception by users. *Frontiers in Medicine*, *5*(SEP). https://doi.org/10.3389/fmed.2018.00260

Allen, H., Cury, A., Gaston, T., Graf, C., Wakley, H., Willis, M., Allen, H., Cury, A., Gaston, T., Graf, C., Wakley, H., & Willis, M. (2019). What does better peer review look like? Underlying principles and recommendations for better practice. *Learned Publishing, 32*(2), 163–175. https://doi.org/10.1002/leap.1222

Annesley, T. M. (2013). Writing an effective manuscript review: The 6 "Be's" to success. *Clinical Chemistry, 59*(7), 1028–1035. https://doi.org/10.1373/clinchem.2013.208280

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality*. https://link.springer.com/book/https://doi.org/10.1007/978-3-319-24106-7

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., & Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education, 31*(2), 145–152. https://doi.org/10.1152/advan.00104.2006

Bharti, P. K., Agarwal, M., & Ekbal, A. (2024). Please be polite to your peers: A multi-task model for assessing the tone and objectivity of critiques of peer review comments. *Scientometrics, 129*(3), 1377–1413. https://doi.org/10.1007/s11192-024-04938-z

Bharti, P. K., Navlakha, M., Agarwal, M., & Ekbal, A. (2023). PolitePEER: Does peer review hurt? A dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation, 58*(4), 1291–1313. https://doi.org/10.1007/s10579-023-09662-3

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*(85), 1–16. https://doi.org/10.1186/s12916-016-0631-5

Cummings, P., & Rivara, F. P. (2002). Reviewing manuscripts for Archives of Pediatrics & Adolescent Medicine. *Archives of Pediatrics & Adolescent Medicine, 156*(1), 11–13. https://doi.org/10.1001/archpedi.156.1.11

Dalkey, N., & Helmer, O. (1963). An Experimental Application of the DELPHI Method to the Use of Experts.Management Science, 9(3), 458–467. https://doi.org/10.1287/mnsc.9.3.458

Delbecq, A. L., Van de Ven, A. H., & G., D. H. G. (1976). Group techniques for program planning: A guide to Nominal Group and Delphi Processes. *Group & Organization Studies, 1*(2), 256–256. https://doi.org/10.1177/105960117600100220

DeMaria, A. N. (2003). What constitutes a great review? *Journal of the American College of Cardiology, 42*(7), 1314–1315. https://doi.org/10.1016/j.jacc.2003.08.020

DiDomenico, R. J., Baker, W. L., & Haines, S. T. (2017). Improving peer review: What reviewers can do. *American Journal of Health-System Pharmacy, 74*(24), 2080–2084. https://doi.org/10.2146/ajhp170190

Fink-Hafner, D., Dagen, T., Doušak, M., Novak, M., & Hafner-Fink, M. (2019). Delphi method: Strengths and weaknesses. *Advances in Methodology and Statistics*. https://doi.org/10.51936/fcfm6982

Ge, M., & Helfert, M. (2007). A review of information quality research-develop a research agenda. *Proceedings of the 2007 International Conference on Information Quality, ICIQ 2007*.

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. In L. Jiao (Ed.), *PLoS ONE* (Vol. 17, Issue 1 January). Public Library of Science. https://doi.org/10.1371/journal.pone.0259238

Giannarou, L., & Zervas, E. (2014). Using Delphi technique to build consensus in practice. *Journal of Business Science and Applied Management*, *9*(2).

Golan, R., Reddy, R., Deebel, N. A., Ramasamy, R., & Harris, A. M. (2023). Peer review: A process primed for quality improvement? *Journal of Urology, 209*(6), 1069–1070. https://doi.org/10.1097/JU.0000000000003460

Goodman, C. M. (1987). The Delphi technique: A critique. *Journal of Advanced Nursing, 12*(6), 729–734. https://doi.org/10.1111/j.1365-2648.1987.tb01376.x

Hill, K. Q., & Fowles, J. (1975). The methodological worth of the Delphi forecasting technique. TechnologicalForecasting and Social Change, 7(2), 179–192. https://doi.org/10.1016/0040-1625(75)90057-8

Horbach, S. P., & Halffman, W. (2018). The changing forms and expectations of peer review. *Research Integrity and Peer Review, 3*(1), 8. https://doi.org/10.1186/s41073-018-0051-5

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation, 12*(10), 1–8.

Jaarsma, T., Strömberg, A., Ärestedt, K., Broström, A., Kärner, A., Mårtensson, J., Moons, P., Thylén, I., & Thompson, D. R. (2013). A good manuscript review for the European Journal of Cardiovascular Nursing. *European Journal of Cardiovascular Nursing, 12*(2), 102–103. https://doi.org/10.1177/1474515113476605

Jamali, H. R., Nicholas, D., Watkinson, A., Abrizah, A., Rodríguez-Bravo, B., Boukacem-Zeghmouri, C., Xu, J., Polezhaeva, T., Herman, E., & Świgon, M. (2020). Early career researchers and their authorship and peer review beliefs and practices: An international study. *Learned Publishing, 33*(2), 142–152. https://doi.org/10.1002/LEAP.1283

Juran, J. M. (1992). *Juran on quality by design: the new steps for planning quality into goods*. Simon and Schuster.

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing, 53*(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

Lee, A. S. (1995). Reviewing a manuscript for publication. *Journal of Operations Management, 13*(1), 87–92. https://doi.org/10.1177/1080569906287960

Liu, C., Cui, J., Shang, R., Jia, Q., Rashid, P., & Gehringer, E. (2024). Generative AI for Peer Assessment Helpfulness Evaluation. *Proceedings of the 17th International Conference on Educational Data Mining*, 412–419. 10.5281/ZENODO.12729848

Lucas, P. J., Baird, J., Arai, L., Law, C., & Roberts, H. M. (2007). Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. *BMC Medical Research Methodology, 7*(1), 4. https://doi.org/10.1186/1471-2288-7-4

Mattioli, J., Robic, P.-O., & Jesson, E. (2022). Information quality: The cornerstone for AI-based Industry 4.0. *Procedia Computer Science, 201*, 453–460. https://doi.org/10.1016/j.procs.2022.03.059

Mitchell, V. -W, & McGoldrick, P. J. (1994). The Role of Geodemographics in Segmenting and Targeting ConsumerMarkets: A Delphi Study. European Journal of Marketing, *28*(5), 54–72. https://doi.org/10.1108/03090569410062032

Niederberger, M., Schifano, J., Deckert, S., Hirt, J., Homberg, A., Köberich, S., Kuhn, R., Rommel, A., Sonnberger, M., Alam, M., Backman, C., Banno, M., Bartoszko, J., Bloomfield, F., Bober, M. B., Chalkoo, M., Chan, T. M., Chen, Y., Coscia, C., Zhang, X. (2024). Delphi studies in social and health

sciences—Recommendations for an interdisciplinary standardized reporting (DELPHISTAR). Results of a Delphi study. *PLOS ONE*, *19*(8). https://doi.org/10.1371/JOURNAL.PONE.0304651

Nils Reimers, I. G. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT. *9th International Joint Conference on Natural Language Processing*, 3982–3992.

Parker, G., Kastner, M., Born, K., & Berta, W. (2021). Development of an implementation process model: A Delphi study. *BMC Health Services Research, 21*(1), 1–12. https://doi.org/10.1186/S12913-021-06501-5

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education, 43*(12), 2263–2278. https://doi.org/10.1080/03075079.2017.1320374

Phillips, J. S. (2011). Expert bias in peer review. *Current Medical Research and Opinion, 27*(12), 2229–2233. https://doi.org/10.1185/03007995.2011.624090

Provenzale, J. M., & Stanley, R. J. (2006). A systematic guide to reviewing a manuscript. *Journal of Nuclear Medicine Technology, 34*(2), 92–99.

Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2013). On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics, 97*(2), 317–356. https://doi.org/10.1007/s11192-013-1002-z

Rai, A. (2016). Editor's comments: Writing a virtuous review. *MIS Quarterly, 40*(3), iii–x.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using Natural Language Processing techniques. *International Journal of Artificial Intelligence in Education, 27*(3), 534–581. https://doi.org/10.1007/s40593-016-0132-x

Ross-Hellauer, T., Deppe, A., & Schmidt, B. (2017). Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PLoS ONE, 12*(12), 1–28. https://doi.org/10.1371/journal.pone.0189311

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15*(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7

Sahil Sanjeev Salvi, S. S. K. (2020). Quality Assurance and Quality Control for Project Effectiveness in Construction and Management . *International Journal of Engineering Research & Technology*, *9*(02).

Severin, A., Strinzel, M., Egger, M., Barros, T., Sokolov, A., Mouatt, J. V., & Müller, S. (2022). *Journal Impact Factor and Peer Review Thoroughness and Helpfulness: A Supervised Machine Learning Study*. https://doi.org/10.48550/arxiv.2207.09821

Shankar, G., & Watts, S. (2003). a Relevant , Believable Approach for Data Quality Assessment. *8th International Conference on Information Quality*, 178–189.

Sizo, A., Lino, A., Reis, L. P., & Rocha, Á. (2018, July 20). An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2018.07.002

Sizo, A., Lino, A., Rocha, Á., & Reis, L. P. (2025). Defining quality in peer review reports: A scoping review. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-025-02435-0

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), 48. https://doi.org/10.1186/s12874-019-0688-x

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*. https://doi.org/10.1136/bmjopen-2019-035604

Thelwall, M. (2020). Automatically detecting open academic review praise and criticism. *Online Information Review, 44*(5), 1057–1076. https://doi.org/10.1108/OIR-11-2019-0347

Venne, V. (2015). Reviewing manuscripts for the Journal of Genetic Counseling: Practical suggestions. *Journal of Genetic Counseling, 24*(2), 189–192. https://doi.org/10.1007/s10897-014-9802-8

Vernon, W., & Vernon, W. (2013). The Delphi technique: A review. *International Journal of Therapy and Rehabilitation, 16*(2), 69–76. https://doi.org/10.12968/ijtr.2009.16.2.38892

Vogel, C., Zwolinsky, S., Griffiths, C., Hobbs, M., Henderson, E., & Wilkins, E. (2019). A Delphi study to build consensus on the definition and use of big data in obesity research. *International Journal of Obesity, 43*(12), 2573–2586. https://doi.org/10.1038/s41366-018-0313-9

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting and Social Change, 79*(8), 1525–1536. https://doi.org/10.1016/J.TECHFORE.2012.04.013

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, J. (2018). Making a difference through quality manuscript review. *Human Resource Development Review, 17*(4), 339–348. https://doi.org/10.1177/1534484318809724

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–34. https://doi.org/10.1080/07421222.1996.11518099

Warner, T. A. (2019). How to write an effective peer-review report: An editor's perspective. *International Journal of Remote Sensing, 40*(13), 4871–4875. https://doi.org/10.1080/01431161.2019.1596342

Wilcox, C. (2019). Rude reviews are pervasive and sometimes harmful, study finds. *Science, 366*(6472), 1433–1433. https://doi.org/10.1126/science.366.6472.1433

Xiong, W., & Litman, D. (2011). Automatically predicting peer-review helpfulness. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *2*, 502–507. http://www.wjh.harvard.edu/

Zhang, J., Zhang, H., Deng, Z., & Roth, D. (2022). *Investigating Fairness Disparities in Peer Review: A Language Model Enhanced Approach*. https://arxiv.org/abs/2211.06398v1

Zhu, Z., Bernhard, D., & Gurevych, I. (2009). Amulti-dimensional model for assessing the quality of answers in social Q&A sites. *Proceedings of the 2009 International Conference on Information Quality, ICIQ 2009*. http://www.ukp.tu-darmstadt.de