Scientometrics (2024) 129:4721–4740
https://doi.org/10.1007/s11192-024-05103-2

Are reviewer scores consistent with citations?

Weixi Xie[1] · Pengfei Jia[1] · Guangyao Zhang[1,2] · Xianwen Wang[1]

Received: 14 July 2023 / Accepted: 1 July 2024 / Published online: 13 July 2024

Abstract
Academic evaluation is a critical component of research, with the interaction between quantitative and qualitative assessments becoming a prominent area of focus. This study examines the relationship between peer review scores and citations within the framework of open peer review. Utilizing data from the OpenReview platform for papers presented at the International Conference on Learning Representations (ICLR), the papers were classified into oral presentations, poster presentations, and rejected manuscripts. Weighted scores were calculated using the confidence score method, followed by an analysis using correlation and regression techniques. The findings reveal significant differences among the three categories in terms of review scores and citations, demonstrating a positive correlation between review scores and citations. Additionally, it was found that papers with greater inconsistency in reviews tended to receive higher citations. Reviewers of rejected papers displayed significantly higher confidence in their assessments compared to reviewers of accepted papers. The results highlight the alignment between peer review and traditional bibliometric indicators in the context of open peer review. However, the degree of concordance between the two evaluation methods is not substantial, suggesting that they are not interchangeable. Therefore, traditional bibliometric indicators should be considered an essential complement to peer review. Furthermore, when evaluating the consistency between quantitative and qualitative assessments and the confidence levels of reviewers, peer review demonstrates greater effectiveness than "traditional peer review" in addressing issues of "poor selection".

Keywords Reviewer score · ICLR · OpenReview · Citation · Open peer review

Introduction

The evaluation criteria of academic research are an ongoing topic of discussion in academia, and one of the key topics is the relationship between qualitative evaluation based on peer review and quantitative indicators. Peer review has long served as the cornerstone of scholarly

✉ Xianwen Wang
xianwenwang@dlut.edu.cn

[1] WISE Lab, Institute of Science of Science and S&T Management, Dalian University of Technology, Dalian, China

[2] UNU-MERIT, Maastricht University, Maastricht, The Netherlands

communication and plays a crucial role in the evaluation of scientific research (da Silva, 2018). Peer review is the process by which scientific journals evaluate and screen submitted articles to ensure their quality before being published. Expert peers are invited to review and provide feedback on submitted articles, and their assessments serve as the primary basis for determining the publication suitability of an article. Another commonly employed method in research evaluation is citation analysis. Citations are the act of authors selecting research materials such as theories, ideas, data, and methods to support their scholarly research. Citation analysis is to use the interdependence of citations and scholarly results for academic evaluation.

Peer review and citation analysis each have advantages and disadvantages in academic evaluation (Derrick & Pavone, 2013). As the gatekeeper of science, peer review plays a vital role in selecting high-quality scientific researchs and providing valuable review comments. However, the pressure of the "publish or perish" culture has led to instances of academic misconduct bypassing the peer review process (Stebbing & Sanders, 2018). Nevertheless, peer review is considered the most reliable method for assessing the quality of academic research in the current academic community (Derrick et al., 2011; Van Raan, 2005). Citation-based metrics are one of the most commonly used metrics to evaluate the influence of scientific publications. It is considered "unobtrusive measures that do not require the cooperation of a respondent and do not themselves contaminate the response" (Smith, 1981). However, citation analysis has been controversial due to the complexity of citation motivations, and the use of citations varies widely across disciplines (Bologna et al., 2022). With standardized applications, citations are considered to be a valid indicator for assessing the impact of scholarship, but it cannot be the only indicator (Aksnes et al., 2019). For most research-oriented scholars, the accuracy and reliability of peer review in research evaluation hold greater credibility. However, some administrators and policymakers may prefer simple and efficient metrics-based evaluation methods (Donovan, 2007). Therefore, our focus is on exploring whether the two evaluation methods can be mutually substitutive by studying the relationship between peer review and citation.

As the open science movement advances, researchers now have the opportunity to empirically investigate the relationship between peer review and bibliometrics, but research is still limited by the openness of peer-reviewed data. In the context of open peer review, this study aims to analyze the variations in review scores and citations of different types of conference papers in computing, as well as the correlation between peer review scores and paper citations, so as to analyze the relationship between peer review results and traditional bibliometric indicators in scientific research evaluation. This provides some references for improving the scientificity of academic evaluation and improving the academic evaluation system.

The rest of the article is organized as follows. Sect. "Literature review" presents the literature review. Sect. "Data and method" then summarizes the data collected to conduct our study and describes the methodology employed to provide our results. Next, Sect. "Results" provides a detailed analysis of these results, while Sect. "Conclusions and final discussion" summarizes our conclusions and provides some final remarks.

Literature review

In this section, given that open peer review constitutes a critical contextual backdrop for this study, we start by providing an overview of open peer review. Subsequently, aligned with the research scope of this article, we undertake a retrospective examination of the literature pertinent to the relationship between peer review and bibliographic indicators.

Open peer review

Peer review dates back to the seventeenth century and first appeared in the Philosophical Repertory of the Royal Society (Bornmann, 2011; Kronick, 1990; Spier, 2002), being the backbone of science. As a screening mechanism, it has been playing an irreplaceable role in the quality control and scientific evaluation of academic journals (Benda & Engels, 2011). Traditional peer review hinges on some degree of anonymity, either single or double blind. Therefore, its implementation encounters several challenges (Seeber & Bacchelli, 2017; Shatz, 2004), such as inadequate review, slow process, rejection of innovative results, lack of trust between reviewers and authors, and unfair assessing based on personal interests and preferences (Fletcher, 1994; Gillespie Jr et al., 1985; Jubb, 2016; Lloyd, 1990; Rennie, 2016). The limitations of peer review have become increasingly noticeable, particularly in the context of the rising importance of interdisciplinary and emerging fields (Bromham et al., 2016; Langfeldt, 2006).

With the continuous advancement of the Open Science movement (Zhang et al., 2021), the Open Peer Review (OPR) has gained increasing popularity in academic journals worldwide, including *PLOS ONE*, *PeerJ*, *BMJ*, and others, due to its emphasis on fairness and transparency. OPR discloses review information to the public, including reviewer and author identities, reviewer recommendations, author responses, and review outcomes, thereby enhancing the transparency, fairness, and efficiency of the review process. However, there is currently no standardized academic definition of open peer review, nor is there a consistent model for its characteristics and implementation. A systematic review has been conducted to define "open peer review," leading to the proposal of a practical definition (Ross-Hellauer, 2017). Nonetheless, the advantages and significance of OPR are widely recognized. The openness of the review process encourages reviewers to be more cautious and fair, improving the quality and objectivity of review opinions while shortening review times. It also enhances the evaluation process monitoring mechanism and promotes knowledge exchange (Nicholson & Alperin, 2016; Zong et al., 2020a, 2020b). However, some studies have found that public review may lead to higher rejection rates and longer review writing times (Van Rooyen et al., 2010). On the other hand, the availability of open peer review reports provides valuable academic information, contributing to a deeper understanding of the peer review process (Zhang et al., 2022). Zong et al., (2020a, 2020b) and Ni et al. (2021) separately examined peer review data from *PeerJ* and *Nature Communications* to investigate the impact of open peer review on manuscript citations, but reached inconsistent conclusions. The former concluded that open peer review increased the citation counts of publications, while the later found no such effect.

The relationship between peer review and bibliometric indicators

Notably, much attention has been paid to the relationship between peer review and traditional bibliometric indicators, especially citations. The most common way to test the validity of peer review is by assessing the impact of publications, and a very common way to do this is based on their citations. Bornmann et al., (2010a, 2010b) collected 1111 publications on *Atmospheric Chemistry and Physics* (ACP), and extracted their citations three years after publication. The results confirm that reviewers' ratings and ACP editorial decisions are related to citations (Bornmann et al., 2010a, 2010b). In a survey of ten computing conference proceedings, Ragone et al. (2011) discovered positive yet weak correlations

between peer review ratings and citations, and found that the peer review process exhibited a high level of randomness. Using 2220 papers submitted to ICLR, Xie et al. (2022) unveiled a weak correlation between conference paper review scores and subsequent citations. Moreover, Tran et al. (2020) observed strong institutional bias and gender differences in the review process. However, using papers from NeurIPS (Conference and Workshop on Neural Information Processing Systems) as the subject of the study, no correlation was found between the evaluation results of accepted conference papers and citations, but correlation existed between rejected papers (Cortes & Lawrence, 2021). Mryglod et al. (2013) analyzed correlations between citations and peer-reviewed results across multiple disciplines, using research groups rather than individuals, and found weak correlations at specific levels across all disciplines. In theory, peer review serves as a filter to select the best research for publication in prestigious journals. However, it has been observed that articles published in lower-level journals often receive higher citation counts, while articles in top journals may receive fewer citations (Jubb, 2016; Smolinsky et al., 2021). Predicting citations is regarded as an essential factor for automatically assessing the future impact of academic publications. In this regard, Li et al. (2019)developed a model to predict publication citations based on reviews.

Due to the availability of data, the relationship between post peer review and citation-based metrics has attracted more attention of researchers. Several studies have demonstrated a significant correlation between publication scores in F1000 and their citations (Li & Thelwall, 2012; Waltman & Costas, 2014). Bornmann (2015) examined the correlation between F1000 expert recommended ratings[1] and traditional literature measures and found a significant positive correlation between FFa and citations. Further, some studies have identified a significant positive correlation between publication scores and citations (Eyre-Walker & Stoletzki, 2013; Mohammadi & Thelwall, 2013). Bornmann further conducted a regression analysis, which showed that the peer review results had a significant effect on the citations. The higher the score of the publications, the higher the number of citations (Bornmann & Haunschild, 2015; Bornmann & Leydesdorff, 2015). In an other work, Zong et al., (2020a, 2020b) investigated the relationship between the polarity of peer review and post-publication citations, finding that publications receiving positive post-publication peer reviews (PPPR's) had significantly higher citation counts when compared to those not receiving positive PPPR's. Smith et al. (2019) conducted multivariable linear regression based on 2361 journal articles in *Gastrointestinal Endoscopy*, and found that the number of F1000 comments was significantly correlated with citations.

The research on the relationship between peer review and other bibliometric indicators has also been the focus of scholarly attention. An analysis of the evaluation outcomes of 56 research projects in the Netherlands revealed varying degrees of correlation between different bibliometric measures and peer review results. Notably, the correlation was found to be higher in the basic sciences compared to the applied sciences (Rinia et al., 1998). In a case study of a Norwegian research group, Aksnes and Taxt (2004) observed a positive, but relatively weak correlation, between all bibliometric indicators and peer review ratings. Overall, most empirical studies in the existing literature demonstrate a positive correlation between peer review outcomes and traditional bibliometric indicators based on citations, with the majority indicating a low correlation coefficient between the two (Jirschitzka et al., 2017).

[1] F1000 is a post-publication peer review system of the biomedical literature. Papers are selected by a peer-nominated leading scientists and clinicians who then rate them (FFa) and explain their importance.

In the context of traditional peer review, the limited availability of data on the review process has hindered the advancement of empirical research on this very important task. Most of the existing studies mentioned above have been conducted in the context of traditional peer review and have reached similar conclusions. Nevertheless, limited research has explored the relationship between peer review and traditional bibliometric indicators within the context of open peer review. The growing adoption of open peer review has provided researchers with ample access to extensive datasets on this specific task, opening up new possibilities for investigation. Drawing upon the data of ICLR conference papers, this study delves into the examination of the interplay between peer review scores and citations within the context of open peer review. Additionally, it undertakes an initial exploratory investigation into the degree of non-consensus and the confidence levels of reviewers. In comparison to prior research, this study offers several notable advantages. Firstly, it utilizes open peer review data obtained from ICLR papers, enabling the quantitative measurement of peer review outcomes through the use of peer review scores. This approach provides an improvement over the predominantly qualitative nature of previous studies. Furthermore, the dataset employed in this study includes rejected papers, thereby enhancing the richness and comprehensiveness of the analysis. Additionally, this study not only investigates the relationship between peer review and bibliometric indicators but also explores the association between the consistency of multiple reviewers' scores within a paper and citations. Due to the strength of the data, we also analyzed the degree of confidence of the reviewers in the peer review process and results.

It is necessary to emphasize that, in comparison to the author's previous research, this article not only utilizes a larger dataset but also incorporates the confidence level of peer reviewers into the calculation of review scores. The results also demonstrate, to a certain degree, the significance of considering the confidence level of peer reviewers in the peer review process. Furthermore, robustness tests were conducted on the regression results in the analysis. Additionally, an independent analysis of the confidence level of peer reviewers was performed in this study. That is to say, this article conducts a deeper and more comprehensive analysis based on the previous research.

Data and method

This study is based on data from ICLR, the International Conference on Learning Representation, which is one of the top conferences in the field of deep learning that employs an open peer review mechanism since 2013. Its data was obtained from a conference paper open review platform called OpenReview available at https://openreviw.net. On this platform, we acquired data on conference papers, including their basic information, review comments from 2–5 expert reviewers, author responses, reviewer ratings of the papers, and the reviewers' confidence levels in their own reviews. The open peer review process of the ICLR conference is illustrated in Fig. 1.

During the review process of ICLR papers, the session chair assumes the responsibility of making the decision to accept or reject a contribution. The session chair considers multiple sources of information, including reviewers' ratings, evidence presented during the review process, discussions between authors and reviewers, and their own evaluation of the papers. Various empirical studies have examined the reliability of this dataset in their research endeavors. For instance, one study analyzed the sentiment of comments and investigated institutional bias in ICLR's comment text data. The findings revealed the presence

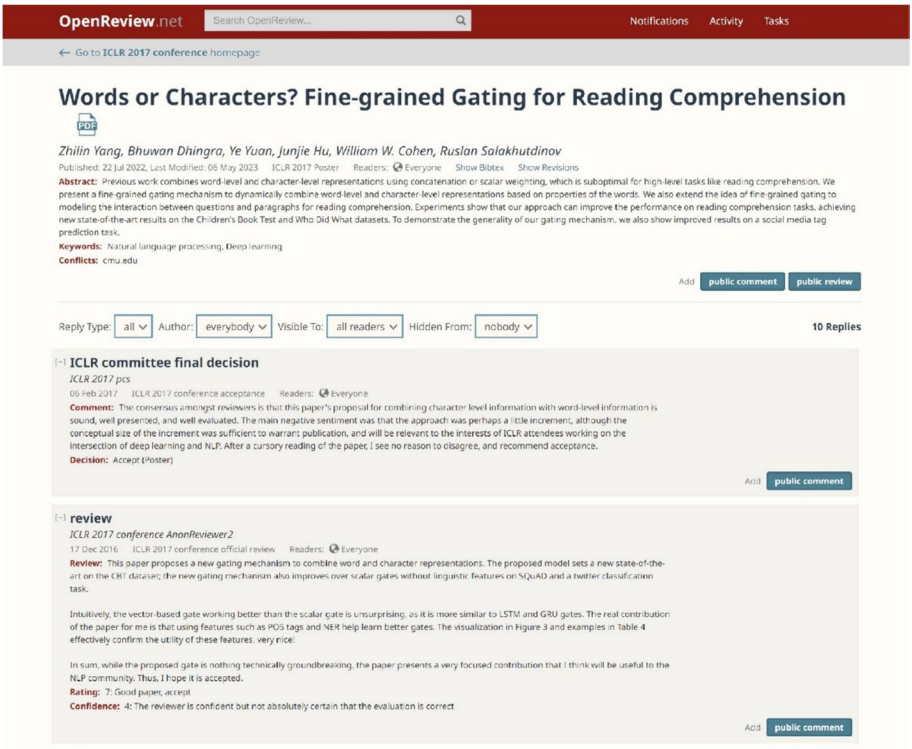

Fig. 1 Schematic diagram of ICLR public review comments

of significant institutional bias in paper selection, with male authors generally obtaining more favorable peer review results and achieving higher citation rates compared to their female counterparts (Tran et al., 2020). Another study employed a multi-instance learning network model to identify the sentiment expressed in review comments based on ICLR 2017 and 2018 reviews, with the aim of predicting the final review scores of papers (Wang & Wan, 2018).

The current study focuses on a comprehensive analysis of scholarly works presented at ICLR conferences between 2017 and 2020, drawn from the corpus of 5352 papers hosted on the OpenReview platform. Among these, 15 papers were excluded from our investigation due to either the absence of review comments or the presence of anomalous data unsuitable for analytical purposes. Consequently, our analysis was conducted on a dataset comprising 5329 papers. Notably, eight papers were untraceable via Google Scholar, leaving us with 5329 viable records for examination. Within this dataset, we identified 110 oral presentations (OPs), 1505 poster presentations (PPs), and 3714 rejected papers (RPs). Our analytical framework was built upon several key variables, including standardized reviewer ratings obtained during the peer review process, reviewers' confidence levels in their assessments, the variance of scores indicating the consistency of individual papers, and the total number of citations garnered by each paper in Google Scholar since its publication (based on a query conducted between June and July 2021). Given the constraints associated with relying solely on conference paper databases, we opted to utilize Google Scholar

citations as our primary source of citation data for this study. It is pertinent to highlight that the chosen analysis period spans from 2017 to 2020, primarily due to the availability of comprehensive peer review data and the assurance of a 2-year citation window for citation counts.

One possible way to measure the certainty (or uncertainty) of one's knowledge is the confidence weighting technique. The degree of a person's knowledge (the cognitive dimension in traditional scoring procedures) and the degree of certainty of his own knowledge can be assessed using confidence-weighted scoring techniques. Confidence-based testing methods were originally proposed by researchers in the fields of education and psychometrics. Although limited studies have been conducted on confidence-weighted scoring methods, consistent findings suggest that these methods yield increased reliability, with no significant differences observed in test validity between weighted and traditional scoring methods (Nathaniel et al., 2021; Rothman, 1969; Zeleznik et al., 1988). In the case of the ICLR data, conference reviewers were asked to indicate their level of confidence in the scores assigned to the papers using a five-point subscale (5, 4, 3, 2, and 1). Confidence weighting techniques (Hausman et al., 1990; Nickerson & McGoldrick Jr, 1965) were employed, with negative weights assigned to scores 1–5 and positive weights assigned to scores 6–10, based on the indicated level of certainty. We calculated the confidence weighted score according to Eq. 1. In this study, we first calculated weighted scores based on the reviewer scores of the papers and the reviewers' confidence in their scores, and used confidence scoring methods. Descriptive statistics were then employed to characterize the dataset, spearman correlation analysis was conducted to explore the relationship between paper revisions, and multiple regression models were utilized to uncover the associations between reviewer scores, paper citations, and the degree of revisions.

$$R_{\mathrm{C}} = R \cdot C \cdot W \tag{1}$$

R_{C} represents the weighted review scores, R represents the review scores given by the reviewers in the original data, C represents the confidence of the reviewers in the review scores they gave, and W represents the weighting factor, where W is negative when R is 1–5 and positive when R is 6–10.

Results

Comparison of indicators in OP, PP and RP

To begin with, this study conducts a comparative analysis of the metrics associated with oral presentations (OP), poster presentations (PP), and rejected papers (RP). Specifically, the review scores and citation counts of the papers since their publication were selected for comparison, as presented in Table 1. The results indicate that OPs exhibited significantly higher review scores and citations compared to PPs, while PPs demonstrated significantly higher review scores and citations compared to RPs. To quantify the level of randomness, one-way ANOVA was employed, a classical method in this regard. The results of the ANOVA revealed statistically significant differences in the mean scores and mean citation counts among the different paper types. Given that the data distribution deviated from normality, the K-S test was subsequently employed for further analysis. A p-value of 0.000 indicated significant differences in the review scores and citations among OPs, PPs, and RPs. As shown in Fig. 2, the comparison of these simple indicators clearly demonstrates

Table 1 Comparison of evaluation indexes of OP, PP and RP

Evaluating indicator	Indicator type	Oral report papers (OP)	Poster presentation papers (PP)	Rejected papers(RP)
Rating	Observations	110	1505	3714
	Average	7.62	6.54	4.98
	Median	7.67	6.67	5
	Variance	0.36	0.37	0.64
Citations	Observations	110	1505	3714
	Average	398.06	123.94	21.69
	Median	221	55	3
	Variance	372161.9	67737.62	4766.43

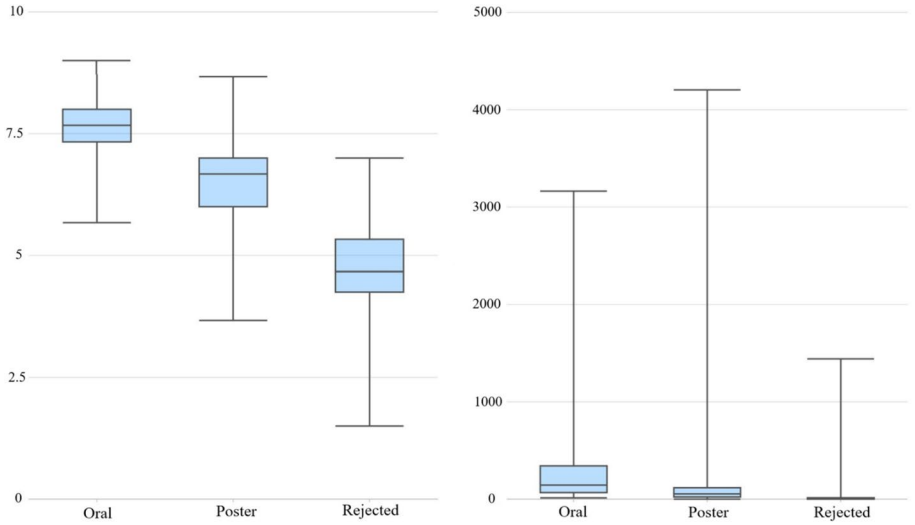

Fig. 2 Distribution of review scores and citations of different types of papers (The left shows the distribution of review scores, and the right shows the distribution of citations)

the substantial distinctions in review scores and citation counts among the three paper categories, thus reflecting the effectiveness of peer review to some extent.

Correlation analysis

In this section, we examine the correlation between paper review scores and citations. In addition, the relationship between the non-consensus degree of a paper and citations is analyzed. Following the K–S test, it was determined that the distribution of review scores and citations did not adhere to a normal distribution. Notably, the distribution of citations exhibited a highly right-skewed and heavy-tailed pattern, with some papers having a substantial number of citations while others had minimal citations. Consequently, the Spearman rank correlation analysis was employed to assess the correlation between various paper metrics, with the Spearman correlation coefficient serving as a measure of the

Table 2 Comparison of evaluation indexes of OP, PP and RP

Category	Spearman correlation coefficient	Observations
ALL	0.714***	5329
OP	0.213***	110
PP	0.197***	1505
OP&PP	0.253***	1615

***Significant correlation when the confidence level (bilateral) is 0.01. Correlation coefficient refers to the correlation between the review score and citations

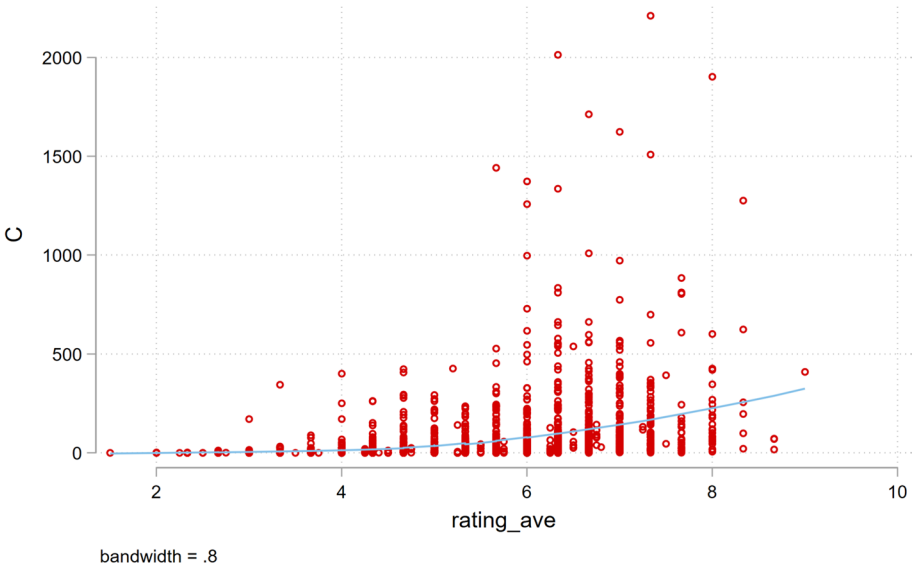

Fig. 3 Scatter chart of review scores (rating_ave) and citations (C) of accepted papers

strength of their nonlinear relationships. As depicted in Table 2, regarding the correlation between review scores and citations, the correlation coefficient for all papers (OP, PP, and RP) was 0.714, indicating a high positive correlation. For OPs, PPs, and accepted papers, the correlation coefficients were 0.213, 0.197, and 0.253, respectively. These coefficients indicate a significant positive correlation between review scores and citations ($p < 0.01$). This result implies that papers receiving higher scores in the peer review process are more likely to be accepted by the conference and receive a greater number of citations. This relationship is also visually evident in Fig. 3.

"Intersubjectivity is equated with realism" (Ziman, 2001). According to Bornmann, scientific discourse is characterized by its pursuit of consensus, as scientific endeavors would be unattainable without scientists reaching similar conclusions (Bornmann, 2015). Some scholars define the reliability of peer review as "the extent to which two or more independent reviews of the same scientific document agree" (Cicchetti, 1991), in other words, the peer review process is considered reliable when there is a high degree of agreement among multiple reviewers. Previous studies have utilized measures such as the non-consensus

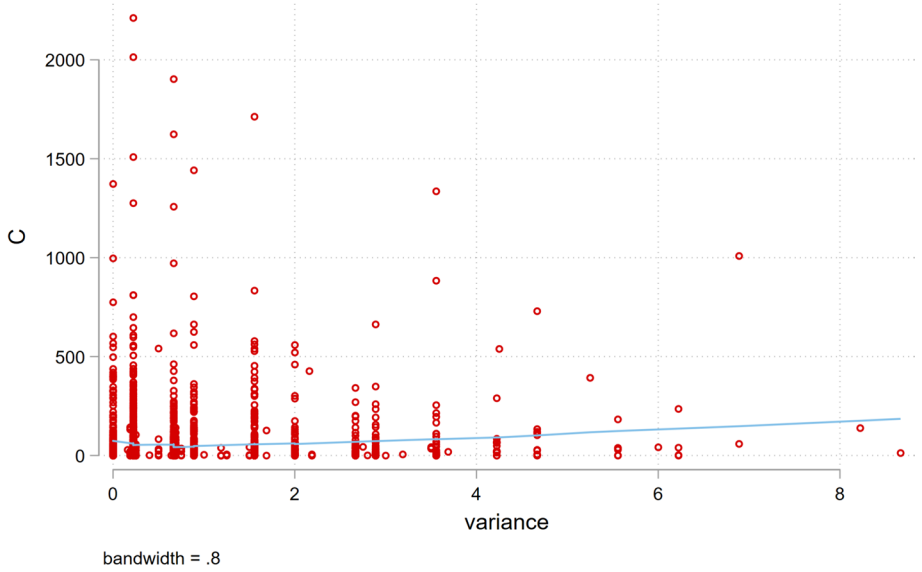

Fig. 4 Scatter chart of the degree of non-consensus (variance) and citations (C) of accepted papers

degree or intraclass correlation coefficient to quantify the level of agreement within peer review groups (Bornmann et al., 2010a, 2010b; Mutz et al., 2012). In this study, we focus on the citation distribution of papers that exhibit discrepancies in reviews, referred to as non-consensus papers. This study considers non-consensus studies as studies that are difficult to reach consensus among multiple reviewers during the review process and are often uncertain and innovative. While a standardized definition of non-consensus research has yet to be established within the academic community, the notion that such research holds value in terms of its innovative contributions has gained widespread recognition. In this study, the variance of paper review scores is employed as an indicator of the overall level of inconsistency within the papers. A larger variance signifies a greater dispersion of inconsistency or a higher degree of disagreement among reviewers.

To visualize the relationship between the two, previous analyses and simple scatter plots indicated that the relationship was not simply linear, so we plotted scatter plots of the relationship between the metrics of the accepted papers using a lowess fit. As shown in Figs. 3 and 4, the relationships between review scores and citations, and between the degree of non-consensus of papers and citations are included. Figure 3 demonstrates a positive correlation between paper review scores and citations; however, the observed correlation is not statistically significant. In addition, we plotted a scatterplot in Fig. 4 based on the degree of non-consensus and citations, and fitting the direction of the curve revealed a positive correlation between the degree of non-consensus and citations.

Regression analysis

In this section, we further investigate the relationship between paper review scores and citations through regression analysis. Descriptive and correlation analyses reveal a positive correlation between paper review scores and citations, as well as a positive correlation

between the degree of reviewer non-consensus and citations. This leads us to inquire whether papers with higher review scores receive more citations. To address these questions, this section employs multiple regression models for in-depth examination.

Based on the aforementioned analysis, this study aims to employ a Negative Binomial Regression (NBR) model to investigate the relationship between peer review scores and citations through regression analysis. The specific model is as follows:

$$\mathrm{Ln}\left(Y_{\mathrm{i}}\right) = \beta X_{\mathrm{i}} + \varepsilon_{\mathrm{i}} \tag{2}$$

Y_{i} represents the citations of the paper, X_{i} is the vector consisting of factors affecting Y_{i}, ε_{i} is the error term. We show the statistics results of variable description and correlation coefficient matrix of variables in Tables 3 and 4, respectively.

Table 5 presents the outcomes of the comprehensive examination of the impact of accepted paper review scores on citation counts. To mitigate the influence of outliers, the values of citations, ratings, and variance were truncated at the 99th percentile prior to regression analysis. Model 1 showcases the regression results incorporating solely the core explanatory variables, with the regression coefficient for review scores exhibiting significant positive association at the 0.01 level. The inclusion of all control variables in model 2 reinforces the significantly positive relationship between review scores and citation counts, while simultaneously enhancing the model's goodness of fit. Given the characteristics of the data, we regress each of the three categories of papers, and models 3–5 show the results obtained from regressing the three categories of papers using a negative binomial regression model. The table shows that the coefficients on the assessment scores are positive for OP and PP and negative for RP. This indicates that the results obtained from regressing the three types of papers separately are consistent, i.e., an increase in paper scores increases the likelihood of a paper's acceptance and leads to a higher number of citations. In addition, the confidence coefficients in models 3 and 4 are negative, whereas the confidence coefficient in model 5 is positive, implying that reviewers show more confidence in evaluating rejected papers than in evaluating accepted papers. In summary, the results of the basic regression analysis suggest that papers that receive higher scores in the peer review process are more likely to receive a higher number of citations.

In this study, robustness tests were conducted employing various methods to validate the consistency of the findings presented in Table 6, thereby further substantiating the hypotheses. First, replacing the explanatory variables in model 6 and using the weighted of ratings to replace the review scores for regression, and model 7 adding control variables on the basis of model 6, it can be found that none of the relevant findings have been substantially changed. In addition, the coefficient of PP is negative and significant at the 0.05 level as seen in model 2 and model 7, which means that OP are cited more often than PP. Rejected papers were further included in the sample for regression, and model 8 and model 9 represent the results of the regression of all papers using raw and weighted rating values, respectively. With all variables included in model 8 and model 9, the regression coefficient for review scores remained positive and significant at the 0.01 level, and the coefficient for accepted papers was significantly positive at the same level. This implies that accepted papers receive higher citation rates compared to rejected papers. Moreover, model 8 and model 9 show that the estimated coefficients of the weighted score values are higher than those of the review scores, and the goodness of fit has improved, which means that the weighted score values taking into account the degree of confidence can explain the relationship between review scores and citations to some extent. In addition, the direction of the variance coefficients in models 8 and

Table 3 Statistical results of variable description ($N = 5329$)

Variable	Type	Definition	Average	Standard deviation	Minimum	Maximum
Citations	Count	Number of citations	139.42	306.47	0.00	4254.00
Rating	Count	Evaluation expert's score for the paper	6.60	0.66	3.67	9.00
Rating_w	Count	Weighted review score	7.09	0.15	1.18	10
Confidence	Count	Confidence level of the reviewer in the paper	3.70	0.54	1.67	5
Variance	Count	Variance of review score	0.85	1.14	0.00	10.89
Accept	Dummy	Accept = 1, Reject = 0	0.37	0.48	0.00	1.00
PP	Dummy	PP = 1, OP = 0	0.94	0.23	0.00	1.00
Reviewer	Count	Number of reviewers	3.13	0.21	1.00	5.00
Year	Dummy	Year of publication	2019.42	0.49	2018.00	2020.00

Table 4 Correlation coefficient matrix of variables ($N=5329$)

Variable	Citations	Rating	Rating_w	Confidence	Variance	Accept	PP	Reviewer	Year
Citations	1								
Rating	0.18	1							
Rating_w	0.09	0.85	1						
Confidence	0.13	0.04	0.23	1					
Variance	0.11	− 0.20	− 0.35	0.16	1				
Accept	–	–	–	–	–	–			
PP	− 0.21	− 0.38	− 0.15	− 0.05	0.00	–	1		
Reviewer	− 0.02	− 0.00	− 0.08	− 0.05	0.00	–	0.02	1	
Year	− 0.21	0.03	0.18	0.01	− 0.09	–	0.04	0.06	1

Table 5 Regression results between paper review scores and citations

	M1	M2	M3	M4	M5
	Accept	Accept	OP	PP	RP
Rating	0.214***	0.278***	0.034*	0.215***	-0.493***
	(0.034)	(0.013)	(0.142)	(0.198)	(0.139)
Confidence		0.102***	− 0.056	− 0.143*	0.217**
		(0.082)	(0.006)	(0.095)	(0.006)
Variance		0.014	0.012	0.015	− 0.073*
		(0.009)	(0.045)	(0.059)	(0.158)
Accept					
PP		− 0.174*			
		(0.018)			
Reviewer		− 0.006	0.012	− 0.018	0.016
		(0.045)	(0.065)	(0.049)	(0.084)
Sample size	1594	1594	109	1485	3672
Year	Y	Y	Y	Y	Y
Adj. R^2	0.17	0.24	0.13	0.38	0.52

Numbers in brackets indicate standard deviation

The standard deviation in brackets ***0.01 **0.05 *0.10, and Y indicates that the time effect is controlled in the model

9 supports the conclusions reached in the previous statistical analysis section. That is, there is a positive correlation between the degree of non-consensus among papers and citations. Model 10 presents the regression results focusing solely on rejected papers, wherein the coefficient of the review score remains positive and significant at the 0.01 level. Additionally, to ensure result independence from the chosen model, model 11 and model 12 adopts poisson regression and ols regression, respectively, revealing that the estimated coefficient of the review score maintains its direction and level of significance without substantial changes. Collectively, the robustness tests conducted utilizing the aforementioned models substantiate the consistency of the estimation results.

Table 6 Robustness test results

	M6	M7	M8	M9	M10	M11	M12
	Accept	Accept	All	All	Rejected	Possion	Ols
Rating			0.516***			0.174**	0.127***
			(0.029)			(0.018)	(0.030)
Rating_w	0.216***	0.196***		0.605***	0.605***		
	(0.034)	(0.072)		(0.005)	(0.006)		
Confidence			0.105**			0.105*	0.116***
			(0.084)			(0.006)	(0.033)
Variance		0.036***	0.038**	0.013***	0.025	0.053	0.030*
		(0.011)	(0.010)	(0.135)	(0.006)	(0.006)	(0.016)
Accept			0.739***	1.053***			
			(0.153)	(0.054)			
PP		− 0.317***	− 0.147**	− 0.319***		− 0.169*	− 0.286**
		(0.091)	(0.009)	(0.034)		(0.058)	(0.082)
Reviewer		0.002	0.014	0.008	0.015	0.008	− 0.013
		(0.031)	(0.031)	(0.012)	(0.138)	(0.106)	(0.090)
Sample size	1594	1594	5266	5266	3672	1594	1594
Year	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.21	0.29	0.52	0.60	0.42	0.35	0.16

Numbers in brackets indicate standard deviation. The standard deviation in brackets ***0.01 **0.05 *0.10, and Y indicates that the time effect is controlled in the model

Analysis of the confidence level of reviewers

In this section, we analyze the level of reviewer confidence in reviewing different types of papers. Peer review has long been recognized as a problem in that it performs poorly in "merit-based ", i.e., it identifies good articles, but hardly identifies outstanding, groundbreaking innovations(Bornmann et al., 2010a, 2010b; Siler et al., 2015). On the contrary, peer review is more effective in "poor selection". Consequently, there have been suggestions to adopt a bottom-line review approach within the peer review process (Brezis, 2007). From the above correlation and regression analyses, we can see that both the consistency of peer review scores with citations and the confidence of reviewers in their own ratings are better for rejected papers compared to accepted papers. Much of the analysis of the consistency of peer review scores with citations has been done in the regression analysis section. Therefore, this section attempts to analyze the confidence level of reviewers for each type of paper. That is, to analyze whether reviewers are more confident in reviewing poor-quality papers (see Table 7).

In order to further explore whether peer review has a better effect of "poor selection " than " merit-based ", an analysis is conducted from the perspective of the confidence of reviewers. The ICLR conference determines by peer review whether an paper is accepted and in what form it is accepted. We therefore analyze the difference in the confidence level of reviewers between the different types of papers based on OP, PP and RP papers. A one-way ANOVA was conducted with the type of paper as the independent variable and the confidence level of the reviewers in rating the paper as the dependent

Table 7 Results of a one-way ANOVA between the type of paper and level of confidence of reviewers

Source	SS	df	MS	*F*	Sig
Between groups	8.082	2	4.041	14.64	0.000
Within groups	612.470	5326	0.276		
Total	620.552	5328	0.279		

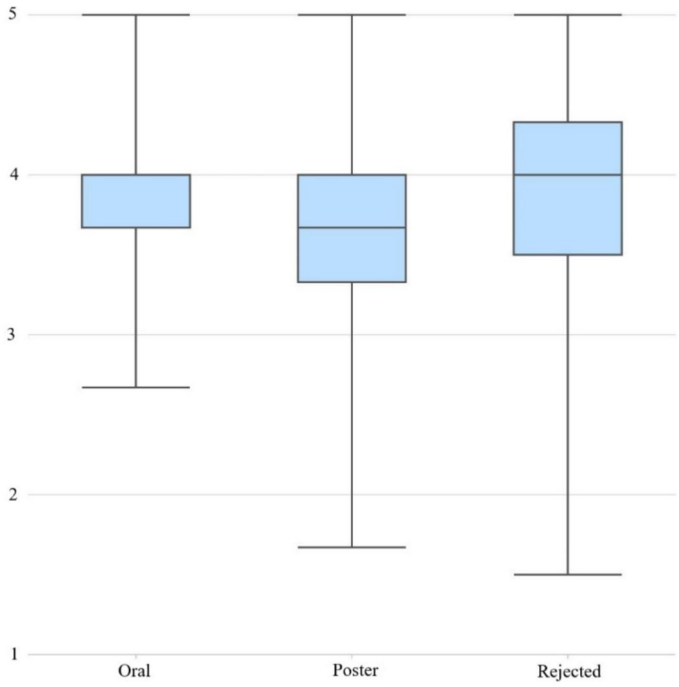

Fig. 5 Distribution of reviewers' confidence level for different types of papers

variable. Table 7 demonstrates the results of the one-way ANOVA. The results showed significant differences ($p < 0.01$) between groups in the confidence levels of the reviewers, which indicates a significant difference between the confidence levels of the reviewers in rating the various types of papers.

Further analyzing the distribution of confidence levels for various types of papers, Fig. 5 demonstrates the confidence levels of reviewers corresponding to different groups of papers. From Fig. 5, it is evident that reviewers exhibited the highest average confidence levels when evaluating rejected papers, whereas the lowest average confidence levels were observed for PPs. This indicates that reviewers demonstrated significantly higher confidence in assessing poorer-quality papers compared to other paper types. However, when it came to evaluating good-quality papers, reviewers displayed a higher degree of uncertainty, particularly in the case of PP papers. In general, we can conclude from the one-way ANOVA in Table 7 and the between-group differences in Fig. 5. That is, when reviewing papers, reviewers have a higher degree of confidence in papers with lower scores, but a lower degree of confidence when reviewing higher-quality papers.

These findings are also supported to some extent by the regression analysis model presented earlier.

Conclusions and final discussion

In this section, we begin in the first subsection by summarizing the findings and drawing conclusions from this study. This is followed by a discussion in the second subsection with the context of the study and a list of limitations of this study.

Conclusions

Peer review has long been the dominant method utilized for assessing scientific research. However, empirical studies have consistently demonstrated that peer review does not consistently yield fair and objective evaluations compared to metric-based assessments (Abramo & D'Angelo, 2011; Ragone et al., 2013). Citations, as a widely adopted bibliometric evaluation metric, serve as indicators of the attention and recognition garnered by research findings from other researchers. Peer review and bibliometric indicators, while distinct evaluation approaches in research assessment, are not opposing methods but rather complementary. The review score of a paper reflects the subjective evaluation by reviewers, which is a qualitative evaluation. Citations, on the other hand, signify the recognition of research by academic peers and to some extent, the quality of research output, representing a quantitative evaluation. Our findings indicate that these two evaluation methods do not always consistently yield congruent results. Based on the peer review data of ICLR papers, descriptive statistics and regression analysis revealed a significant correlation between the review scores of papers and citations. Specifically, higher review scores were associated with increased citations. Moreover, a positive correlation was observed between the degree of discrepancy among reviewers regarding papers and citations. Additionally, it was discovered that reviewers exhibited greater confidence in reviewing poorer-quality papers but were less confident when evaluating higher-quality papers.

ICLR determines the acceptance of papers through open peer review and decides whether a paper is accepted as an oral presentation or a poster presentation. Descriptive statistics, ANOVA and correlation analysis of OP, PP and RP showed that the review scores and citations differed among the three types of papers, with statistically significant differences ($p < 0.05$). The correlation analysis also found that the review score of papers was significantly and positively correlated with citations. Additionally, it was observed that the degree of non-consensus of reviewers for papers was also positively correlated with citations. These statistical outcomes provide evidence supporting the overall validity of the peer review process and its alignment with traditional bibliometric indicators.

This study employed multiple regression models to further investigate the relationship between paper review scores and citations. The results showed that there was a significant positive correlation between review score and citations for PP, accepted papers and all papers. These findings support previous research highlighting the association between peer review results and citations. The results suggest that although peer review and citations provide different perspectives on the academic impact of scientific research, they are positively correlated to some extent. This demonstrates the consistency between peer review and citations in the evaluation of scientific research. Furthermore, the findings reveal a significant correlation between these two indicators of scientific research evaluation. While

peer review and citations are just one aspect of assessing the quality and impact of a paper, their correlation implies that they measure conceptually similar constructs. Additionally, based on the degree of public review in the ICLR dataset, this study explores the citation distribution of non-consensus papers and finds that papers with a higher degree of non-consensus tend to receive more citations.

It was also found that, in terms of the degree of agreement between paper review scores and citation frequency, rejected papers were higher than PP papers than OP papers. This indicates to some extent that the validity of peer review is more pronounced in the lower-quality papers and less in the higher-quality papers. The regression analysis process also uncovered variations in the confidence levels of reviewers when evaluating different types of papers. Therefore, further analysis of reviewers' confidence levels revealed that reviewers held higher levels of confidence in identifying lower-quality papers and lower levels of confidence in the review process of higher-quality papers. This analysis underscores the challenges faced by peer review, as it demonstrates that while peer review is effective in identifying lower-quality papers, it encounters difficulties in accurately evaluating higher-quality papers.

Final discussion

In summary, this study aligns with previous research by demonstrating a positive correlation between peer review scores and paper citations. We contribute to the existing literature by conducting a statistical analysis of open peer review data from ICLR papers, which enhances our understanding of the validity and reliability of peer review. The results of peer review are judged from the perspective of reviewers, while traditional metrics consider the quality and impact of papers from the perspective of authors. As the "gatekeeper" of scientific research, peer review may lead to unfair results due to subjective bias. However, as the primary mechanism to control the quality of scientific research, it plays a crucial role in the construction of the research evaluation system. As the basis of traditional bibliometric evaluation, citations reflects, to some extent, peer evaluation of research quality and impact, although it is incomplete and biased. Our study found that although there is a positive correlation between peer-reviewed results and citation metrics, they are not substitutable relationships for each other. Peer review remains the most crucial component of the current research evaluation system. Compared with the elite evaluation of peer review, bibliometric indicators can provide a broader reference of public peer evaluation. Therefore, a comprehensive scientific research evaluation system should be based on the quality assessment provided by qualitative peer review while integrating bibliometric indicators to form a balanced evaluation model that combines subjective and objective aspects. Furthermore, reviewers' confidence levels during the peer review process can provide valuable insights. Taking into account both the level of confidence and non-consensus among reviewers can assist editors in making informed decisions to some extent.

This study has several limitations that should be acknowledged. Firstly, the utilization of open peer review data in this study offers advantages in terms of transparency and accessibility. However, it is important to note that the open peer review model is not widely adopted by all journals and conferences, and the level of openness can vary. Consequently, the generalizability of the findings may be limited. Secondly, this paper was conducted on conference papers in the field of computing, and caution is needed when inferring conclusions due to possible disciplinary differences. Third, this study analyzes the correlation between peer review results and bibliometric indicators, and proposes that both

qualitative and quantitative evaluation tools should be combined for effective research evaluation. However, how to achieve the integration of the two types of evaluation is an urgent issue that needs further research. Furthermore, it is worth noting that this study was conducted using a single dataset due to data availability. The relatively small sample size of papers analyzed in this study may influence the results to some extent. Addressing these limitations and exploring avenues for integrating qualitative and quantitative evaluation in research assessment should be the focus of future research.

Acknowledgements We gratefully acknowledge the grants received from the National Natural Science Foundation of China (71974029) and from the Social Science Foundation of China (22&ZD194). We extend our sincere gratitude to the reviewers for their insightful comments and constructive feedback, which have significantly enhanced the quality of this manuscript.

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics, 87*(3), 499–514.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open, 9*(1), 2158244019829575.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a norwegian university. *Research Evaluation, 13*(1), 33–41.

Benda, W. G., & Engels, T. C. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting, 27*(1), 166–182.

Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2022). Do open citations give insights on the qualitative peer-review evaluation in research assessments? An analysis of the Italian national scientific qualification. *Scientometrics, 128*, 1–35.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1), 197–245.

Bornmann, L. (2015). Interrater reliability and convergent validity of F 1000 P rime peer review. *Journal of the Association for Information Science and Technology, 66*(12), 2415–2426.

Bornmann, L., & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and mendeley. *Journal of Informetrics, 9*(3), 477–487.

Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. *Journal of Informetrics, 9*(3), 419–429.

Bornmann, L., Marx, W., Schier, H., Thor, A., & Daniel, H.-D. (2010a). From black box to white box at open access journals: Predictive validity of manuscript reviewing and editorial decisions at atmospheric chemistry and physics. *Research Evaluation, 19*(2), 105–118.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010b). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE, 5*(12), e14331.

Brezis, E. S. (2007). Focal randomisation: An optimal mechanism for the evaluation of R&D projects. *Science and Public Policy, 34*(10), 691–698.

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature, 534*(7609), 684–687.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119–135.

Cortes, C., & Lawrence, N. D. (2021). Inconsistency in Conference peer review: revisiting the 2014 neurips experiment. Preprint retrieved from. arXiv:2109.09774

da Silva, J. A. T. (2018). Challenges to open peer review. *Online Information Review, 43*(2), 197–200.

Derrick, G. E., Haynes, A., Chapman, S., & Hall, W. D. (2011). The association between four citation metrics and peer rankings of research influence of Australian researchers in six fields of public health. *PLoS ONE, 6*(4), e18521.

Derrick, G. E., & Pavone, V. (2013). Democratising research evaluation: Achieving greater public engagement with bibliometrics-informed peer review. *Science and Public Policy, 40*(5), 563–575.

Donovan, C. (2007). *Introduction: Future pathways for science policy and research assessment: Metrics vs peer review quality vs impact* (pp. 538–542). Beech Tree Publishing.

Eyre-Walker, A., & Stoletzki, N. (2013). The assessment of science: The relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biology, 11*(10), e1001675.
Fletcher, S. (1994). Guardians of science: Fairness and reliability of peer review. *BMJ, 309*(6952), 488.
Gillespie, G. W., Jr., Chubin, D. E., & Kurzon, G. M. (1985). Experience with NIH peer review: Researchers' cynicism and desire for change. *Science, Technology, & Human Values, 10*(3), 44–54.
Hausman, C. L., Weiss, J. C., Lawrence, J. S., & Zeleznik, C. (1990). Confidence weighted answer technique in a group of pediatric residents. *Medical Teacher, 12*(2), 163–168.
Jirschitzka, J., Oeberst, A., Göllner, R., & Cress, U. (2017). Inter-rater reliability and validity of peer reviews in an interdisciplinary field. *Scientometrics, 113*, 1059–1092.
Jubb, M. (2016). Peer review: The current landscape and future trends. *Learned Publishing, 29*(1), 13–21.
Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. *JAMA, 263*(10), 1321–1322.
Langfeldt, L. (2006). The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation, 15*(1), 31–41.
Li, S., Zhao, W. X., Yin, E. J., & Wen, J.-R. (2019). A neural citation count prediction model based on peer review text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)
Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. Proceedings of the 17th International Conference on Science and Technology Indicators
Lloyd, M. E. (1990). Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis, 23*(4), 539–543.
Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. *Scientometrics, 97*(2), 383–395.
Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics, 97*, 767–777.
Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS ONE, 7*(10), e48509.
Nathaniel, E. D., Scott, H. F., Wathen, B., Schmidt, S. K., Rolison, E., Smith, C., & Lockwood, J. M. (2021). Confidence-weighted testing as an impactful education intervention within a pediatric sepsis quality improvement initiative. *Pediatric Quality & Safety*. https://doi.org/10.1097/pq9.0000000000000460
Ni, J., Zhao, Z., Shao, Y., Liu, S., Li, W., Zhuang, Y., & Li, J. (2021). The influence of opening up peer review on the citations of journal articles. *Scientometrics, 126*, 9393–9404.
Nicholson, J., & Alperin, J. P. (2016). A brief survey on peer review in scholarly communication. *The Winnower*. https://doi.org/10.1371/journal.pone.0189311
Nickerson, R. S., & McGoldrick, C. C., Jr. (1965). Confidence ratings and level of performance on a judgmental task. *Perceptual and Motor Skills, 20*(1), 311–316.
Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2011). A quantitative analysis of peer review. In13th International Society of Scientometrics and Informetrics Conference, Durban
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*, 317–356.
Rennie, D. (2016). Let's make peer review scientific. *Nature, 535*(7610), 31–33.
Rinia, E. J., Van Leeuwen, T. N., Van Vuren, H. G., & Van Raan, A. F. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy, 27*(1), 95–107.
Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research*. https://doi.org/10.12688/f1000research.11369.2
Rothman, A. I. (1969). Confidence testing: An extension of multiple-choice testing 1. *Medical Education, 3*(3), 237–239.
Seeber, M., & Bacchelli, A. (2017). Does single blind peer review hinder newcomers? *Scientometrics, 113*(1), 567–585.
Shatz, D. (2004). *Peer review: A critical inquiry*. Rowman & Littlefield.
Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences, 112*(2), 360–365.
Smith, L. C. (1981). Citation analysis. *Library Trends, 30*(1), 83–106.
Smith, Z. L., Chiang, A. L., Bowman, D., & Wallace, M. B. (2019). Longitudinal relationship between social media activity and article citations in the journal gastrointestinal endoscopy. *Gastrointestinal Endoscopy, 90*(1), 77–83.
Smolinsky, L., Sage, D. S., Lercher, A. J., & Cao, A. (2021). Citations versus expert opinions: Citation analysis of featured reviews of the American mathematical society. *Scientometrics, 126*, 3853–3870.
Spier, R. (2002). The history of the peer-review process. *TRENDS in Biotechnology, 20*(8), 357–358.

Stebbing, J., & Sanders, D. (2018). The importance of being earnest in post-publication review: Scientific fraud and the scourges of anonymity and excuses. *Oncogene, 37*(6), 695–696.
Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., & Goldstein, T. (2020). An open review of openreview: A critical analysis of the machine learning conference review process. Preprint retrieved from arXiv:2010.05137
Van Raan, A. F. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*, 133–143.
Van Rooyen, S., Delamothe, T., & Evans, S. J. (2010). Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *BMJ*. https://doi.org/10.1136/bmj.c5729
Waltman, L., & Costas, R. (2014). F 1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology, 65*(3), 433–445.
Wang, K., & Wan, X. (2018). Sentiment analysis of peer review texts for scholarly papers. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval
Xie, W., Zhang, G., & Wang, X. (2022). Relationship between peer review score and cited frequency of conference papers under the background of open peer review. *Chinese Journal of Scientific and Technical Periodicals, 33*(1), 113–121.
Zeleznik, C., Hojat, M., Goepp, C. E., Amadio, P., Kowlessar, O., & Borenstein, B. (1988). Students' certainty during course test-taking and performance on clerkships and board exams. *Academic Medicine, 63*(12), 881–891.
Zhang, G., Wang, L., Xie, W., Shang, F., Xia, X., Jiang, C., & Wang, X. (2022). "This article is interesting, however": Exploring the language use in the peer review comment of articles published in the BMJ. *Aslib Journal of Information Management, 74*(3), 399–416. https://doi.org/10.1108/AJIM-06-2021-0172
Zhang, G., Wang, Y., Xie, W., Du, H., Jiang, C., & Wang, X. (2021). The open access usage advantage: A temporal and spatial analysis. *Scientometrics, 126*, 6187–6199.
Ziman, J. (2001). *Real science: What it is, and what it means*. IOP Publishing.
Zong, Q., Fan, L., Xie, Y., & Huang, J. (2020a). The relationship of polarity of post-publication peer review to citation count: Evidence from Publons. *Online Information Review, 44*(3), 583–602.
Zong, Q., Xie, Y., & Liang, J. (2020b). Does open peer review improve citation count? Evidence from a propensity score matching analysis of PeerJ. *Scientometrics, 125*(1), 607–623.