Scientometrics
https://doi.org/10.1007/s11192-025-05422-y

Scientific publishing without gatekeeping: an empirical investigation of *eLife's* new peer review process

Rüdiger Mutz[1] · **Lutz Bornmann**[2] · **Hans-Dieter Daniel**[3]

Received: 18 December 2024 / Accepted: 22 August 2025

Abstract
At the end of January 2023, *eLife* introduced a new publishing model (alongside the old-traditional-publishing model): all manuscripts submitted as preprints are peer-reviewed and published if they are deemed worthy of review by the editorial team ("editorial triage"). The model abandons the gatekeeping function and retains the previous "consultative approach to peer review". Even under the changed conditions, the question of the quality of judgements in the peer review process remains. In this study, the reviewers' ratings of manuscripts submitted to *eLife* were examined in terms of both descriptive comparisons of peer review models, and the following selected quality criteria of peer review: interrater agreement and interrater reliability. *eLife* provided us with the data on all manuscripts submitted in 2023 according to the new publishing model (group 3, $N = 3{,}846$), as well as manuscripts submitted according to the old publishing model (group 1: $N = 6{,}592$ submissions from 2019; group 2: $N = 364$ submissions from 2023). The interrater agreement and interrater reliability for the criteria "significance of findings" and "strength of support" were similarly low, as previous empirical studies for gatekeeping journals have shown. The fairness of peer review is not or only slightly compromised. We used the empirical results of our study to recommend several improvements to the new publishing model introduced by *eLife* as for example, increasing transparency, masking author identity or increasing the number of expert reviewers.

Keywords Journal peer review · Interrater agreement · Interrater reliability · Gatekeeping · Open science

✉ Rüdiger Mutz
ruediger.mutz@uzh.ch

[1] Center for Higher Education and Science Studies (CHESS), University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland

[2] Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, 80539 Munich, Germany

[3] University of Zurich, 8050 Zurich, Switzerland

Published online: 01 October 2025

Introduction

At the end of January 2023, the editors and editorial team of *eLife* introduced a new publishing model that breaks with the previous traditional model. Manuscripts submitted as preprints can be published if they are deemed worthy of review by the editorial team of over 70 senior editors and more than 700 review editors (Abbott, 2023; Eisen et al., 2020, 2022; Graham, 2022; Urban et al., 2022). In this new model, "the Reviewing Editors offer their scientific views on the paper, and there is an open discussion about whether to review the paper, using the reasoning laid out above. Editors are considering whether the work is of substantial interest, whether they will be able to find high-quality reviewers, and whether the reviews will be valuable to the scientific community" (eLife Editorial, eLife Senior, & eLife Early Career Advisory, 2024, p. 2). A review of these pre-selected manuscripts ("editorial triage") still takes place, but the final editorial decision to publish is made on a "publish, then review" basis (Eisen et al., 2020, p. 1). The previous "consultative approach to peer review" (Schekman, 2017, p. 1) has been retained. In the new publishing model, authors are given advice on how to revise their manuscripts, and readers are helped to better assess the quality of the published article. Both address the function of peer review to improve an article, which has been considered an important function for many years: "Another answer to the question of what is peer review for is that it is to improve the quality of papers published or research proposals that are funded" (Smith, 2006, p. 179). To this end, *eLife* has introduced two additional peer review criteria as rating scales: "significance of findings" and "strength of support". The review process at *eLife* ends when the authors have decided that a contribution should be published, and the reviews and the ratings are published together with the manuscript. As authors are (still) free to choose whether to use the old or the new publishing model, manuscripts were available for both publishing models in 2023 (for this study).

The decision to introduce a new publishing model was preceded by a long discussion at *eLife*, which was influenced by demands in the spirit of the current Open Science movement. For example, *eLife* rejects the journal impact factor in line with the *San Francisco Declaration on Research Assessment* (DORA, https://sfdora.org) (Marder, 2014; Schekman, 2019).

With regard to the new publishing model, a trial was carried out in 2018 "to test the feasibility of an even more radical form of peer review—an approach in which the authors will control the decision about publication and how they respond to the comments made by peer reviewers" (Patterson & Schekman, 2018, p. 1). With this approach, the name of the journal should no longer serve as a proxy for the potential quality of the contributions. The journal should become the venue for critical discussion of a contribution. Reviewers lose their gatekeeper role and the review process should encourage a constructive and transparent dialogue, as reviews and decisions are made public. The efficiency of research evaluation at journals may be increased by the new approach as manuscripts no longer have to be rejected and submitted to other journals (Patterson & Schekman, 2018). As all peer-reviewed manuscripts are published by *eLife*, the *eLife* assessments have a special significance: "During the review process, editors and reviewers discuss their reviews with each other and assess the significance of the findings and the strength of the evidence reported in the preprint. Their conclusions are captured in an 'eLife assessment', which is written with the help of a common vocabulary to ensure consistency" (https://elifesciences.org/peer-review-process). Readers of papers published in *eLife* are given information about the quality of the research presented.

With the introduction of the new publishing model, *eLife* is taking a radical step forward, both in comparison with the previous model at *eLife* and in comparison with other prominent journals in the field of life sciences and medicine. Most journals continue to rely on the classic gatekeeping procedure. It is therefore not surprising that the move has not been without criticism and resistance (e.g., Abbott, 2023; Else, 2022; Urban et al., 2022). According to Urban et al. (2022) *eLife* relies on "editorial triage": A small number of editors decide whether a preprint will be peer-reviewed and therefore accepted for publication or not. However, such decisions are particularly prone to errors of judgment. Urban et al. (2022) therefore call for an initial selection process that follows transparent rules and is protected against bias and unfairness. According to Abbott (2023), in November 2022, 47 editors wrote a letter to the (then) editor in chief, Michael Eisen, in which they pointed out possible losses in the quality of papers if it were no longer possible to reject papers in the peer-review process. "They worried about harm to the journal's collaborative open-reviewing process, and that the quality of papers on the *eLife* platform would drop. With no possibility of rejection, some authors might choose to ignore reviewer comments or only superficially address them, they wrote—and that knowledge might discourage reviewers from producing detailed critiques" (Abbott, 2023, p. 781). Authors from various countries where hiring and promotion still depend on the name of the journal (in which they publish) would no longer submit their manuscripts to *eLife*. Another criticism at the new publishing model concerns the filter function of journals (Abbott, 2023, p. 781). The model does not allow for transparent and comprehensible filtering upstream of articles, especially for readers outside the scientific community. It is not clear whether an article is generally trustworthy or not.

eLife is not only open to change, but also to empirical studies and experiments that follow the changes. The journal has repeatedly been the subject of empirical studies, e.g., on the question of whether there are qualitative differences in review reports and decision letters before and after the pandemic for Covid-19 related articles and non-Covid-19-related articles (Horbach, 2021). Other studies have looked at "the effects of an editor serving as one of the reviewers during the peer-review process" (Vaggi et al., 2016) or the relationship between reviewers' academic status and their use of language (emotional features or linguistic features) (Sun et al., 2023). A pre-registered study by Hardwicke et al. (2024) was prompted by possible inconsistencies in the two review criteria "significance" and "strength of support". In an online repeated measurement experiment, the authors found a mismatch of 20% or less between the intended ranking of phrases from the *eLife* vocabulary (e.g., landmark, fundamental, important, valuable, useful) and the actual ranking assigned by a sample of graduates with or without a doctoral degree. An alternative vocabulary presented by the authors performed significantly better (60% or more).

eLife itself has also commissioned empirical studies (e.g., Rodgers, 2017). Initial results are also available from surveys of authors, editors, and readers conducted 3 and 6 months after the introduction of the new publishing model (2023b; Ratan et al., 2023a). After 6 months, author surveys showed mostly positive feedback. The most important reason for submitting to *eLife* is the quality of the journal (78.57%), the second most important reason is to try the new model (67.86%). Senior editors "mostly support the new model but have concerns about maintaining quality" (Ratan et al., 2023a). As the *eLife* assessments are published together with the peer-reviewed preprints, readers' perceptions of the assessments are of particular interest: "50% had read at least one and 75% of those readers said they were useful to them" (Ratan et al., 2023a).

A brief state of the art of (open) peer review

The developments at *eLife* can also be seen against the background of peer review research, especially the development of open peer review models (Armstrong, 1997; Ross-Hellauer, 2017; Ross-Hellauer et al., 2023). Open peer review, "despite being a major pillar of Open Science, has neither a standardized definition nor an agreed schema of its features and implementations" (Ross-Hellauer, 2017, p. 1). In a systematic literature review, Ross-Hellauer (2017, p. 7) identifies seven elements of open peer review, three of which are central: Open identities ("Authors and reviewers are aware of each other's identity"), open reports ("Review reports are published alongside the relevant article"), and open participation ("The wider community are able to contribute to the review process"). Open peer review is seen as a response to key problems of classical peer review, such as "delay", i.e. "unacceptable slowing down of the scientific process", a "lack of transparency" and "lack of recognition for reviewers" (Walker & Rocha da Silva, 2015, p. 3f). Furthermore, classical peer review, in particular, is criticised for failing to fulfil quality criteria such as inter-rater reliability, predictive validity, and fairness (Aczel et al., 2025; Bornmann & Daniel, 2008a, 2008b, 2010a, 2010b, 2010c; Cicchetti, 1991; Lee et al., 2013; Nicolai et al., 2015).

Waltman et al. (2023) characterize this line of research focusing on quality criteria of peer review as one of four schools of peer review research ("Quality & Reproducibility School"). One of the few stable findings from this school of peer review research is the low level of agreement between reviewers in their assessment of the same manuscript. Bornmann et al. (2010) showed in a meta-analysis of 48 studies that the interrater reliability and interrater agreement of the journal peer review is low (ICC/$r^2 = 0.34$, mean Cohen's Kappa = 0.17). With regard to grants, Marsh et al., (2008, p. 162) considers interrater reliabilities of at least 0.80/0.90 to be sufficient to recognize differences between two proposals. Although this result may be used as an argument for abandoning final publication decisions, with the introduction of review criteria the questions of sufficient interrater agreement, interrater reliability, validity, and fairness of the peer review process remain. Only if the peer review process ensures the quality of publication decisions, readers can use the criteria to assess an article in the sense of a filter function. There is less evidence yet with respect to open peer review. F1000Prime's post-publication peer review, for example, shows a low level of agreement among faculty members (Bornmann, 2015).

Questions of validity can be answered by correlating with assessments in peer review with external quality criteria such as citations. For example, Cheng et al. (2024) investigated the influence of open peer review on citations and altmetrics in two journals: *Nature Communications* and PLoS One. Open peer review articles were no more likely to be cited than non-open peer review articles in the two journals, but regarding altmetrics, open peer review articles were more cited on average than non-open peer review articles. Since a citation impact measurement requires a citation window of at least 3 years to arrive at reliable citation counts (e.g., Bornmann & Daniel, 2008a), the validity of the peer review process cannot be investigated. Only current data on the new peer review model is available to date.

The fairness of reviewers' assessments can be defined by the absence of bias: "… reviewer bias is understood as the violation of impartiality in the evaluation of a submission. We define impartiality in peer evaluations as the ability for any reviewers to interpret and apply evaluative criteria in the same way in the assessment of a submission"

(Lee et al., 2013, p. 4). The review process is fair if the reviewers' ratings are free of bias, that is the assessments should be independent of characteristics unrelated to the quality of the research presented (e.g., characteristics of authors or reviewers) (Bornmann & Daniel, 2009; Daniel, 1993; Mutz et al., 2015; Squazzoni et al., 2021).

We derived the following three research questions from the specific developments at *eLife* and the requirements of peer review processes to ensure the quality of published research:

1. *Descriptive comparison of peer review model*: What are the differences between the old and new publishing models in terms of submission characteristics?

   In the new publishing model, two review criteria (rating scales) were introduced by *eLife* in the peer review process: "significance of findings" and "strength of support".

2. *Interrater reliability of peer review:* How do reviewers agree in their assessments of manuscripts submitted under the new publishing model with respect to both review criteria?
3. *Fairness of peer review:* Are there any factors that influence peer review ratings that are part of the formal peer review process?

Data and methods

Data

In contrast to a classical empirical study, this study does not collect primary data, but rather re-analyzes data generated as part of a publishing and assessment process. This requires that the data generated are processed in such a way that allows statistical analysis. From *eLife*, datasets were available on the authors, on the submitted manuscripts, on the preprint server of the submitted manuscript (e.g., bioRxiv, medRxiv), on the review outcomes in the old and new publishing model, and on triage rejection letters. The datasets cover the entire process from submission to publication, distinguishing between the old and new publishing model. All submissions that were handled by one or more editors were included in this study. The different datasets were merged using the manuscript ID or author ID. A total of 10,802 documents were included in the statistical analysis. As many submissions were desk rejected, only 2969 submissions were subjected to peer review (27.5% of all submissions).

The focus of the data analysis in this study was on the new publishing model, which was introduced by *eLife* in 2023. To compare the old publishing model with the new model, the data from all manuscripts submitted in 2019 were defined as the comparison group. Submissions from 2019 but not later were selected as the comparison group, as irregularities due to the coronavirus pandemic were to be expected in 2020 and 2021 (e.g., publications or longer peer review). As manuscripts could still be submitted under the old publishing model in 2023, we also included this cohort in the study to have a direct comparison between the old and new publishing models in the same year.

We considered three groups in the study: old publishing model in 2019 (group 1), old publishing model 2023 (group 2), and new publishing model 2023 (group 3). While the dataset for the 2019 publication year is complete, only 2023 submissions from January to

October could be considered due to the timing of data access for this study. The new publishing model started in February 2023. To ensure comparability between the three groups, only manuscripts submitted between February and October 2019 or 2023 were included in the analysis (group 1: $N=6{,}592$ manuscripts, group 2: $N=364$ manuscripts, group 3: $N=3{,}846$ manuscripts). All manuscripts submitted in 2023 under the new publishing model for which peer review ratings were available were included in the reliability analyses. That's only just 875 out of 3,846 manuscripts that were submitted.

Variables

The validity of the data was guaranteed, as far as possible, through the use of summary statistics, plausibility checks, and feedback to the editorial team. With respect to missing values, two types of missing values can be distinguished: systematic and random missing values. Systematic missing values occurred, for instance, due to the lack of data for the old publishing model. For the random missing data, it was assumed that the missings are completely at random (MCAR), i.e., the missing process is completely random (Enders, 2025). There are no variables in the dataset or yet to be included that cause missing values.

Three types of information about the submissions are of interest for answering our research questions: first, information on the documents; secondly, information about the submission process. Both types of information were available on a comparable basis for all three groups studied. Third, information on the peer review ratings was used in the study. As no quantitative scales were used in the old publishing model, the ratings in quantitative form were only available for the new publishing model.

The following information about the *documents* was analyzed:

- Number of appeals per submission, number of words in title, document type (research article, short reports, tools & resources, research advances), duration of the review process.
- *eLife* distinguishes between 18 research areas, which can be roughly divided into two categories "Biology & Biochemistry" and "Clinical Medicine".

The following information was included in the study concerning the *submission process:*

- Submissions: reviewing editors, who are involved in the review process, and senior editors and their country of residence, final decision (rejection, acceptance).
- Decisions: initial rejection ("editorial triage"), final and total rejection (i.e. the total number of initial and final rejections), acceptance. Unlike the new publishing model, manuscripts in the old model have to overcome two hurdles: the initial assessment (e.g., desk reject) and the final assessment.

For the *reviewers' ratings* in the new publishing model (the third type of information considered here), internal assessments of each reviewer on 5- and 6-point rating scales were available for the criteria "significance of finding" and "strength of evidence" criteria. On both scales, an additional 0="prefer not to answer" was included. According to the website (https://elifesciences.org/about/elife-assessments), "Every Reviewed Preprint published in the journal now includes an *eLife* Assessment written by the editor who oversaw the review process and the peer reviewers. This assessment summarizes what the editor

and reviewers thought about the article." According to another website, "Public Reviews describe the article's strengths and weaknesses, and indicate whether its claims and conclusions are supported by the data" (https://elifesciences.org/about/peer-review). The *eLife* Assessment, Public Reviews and names of editors and reviewers are published finally alongside the revised manuscript.

Statistical analyses

Interrater-agreement and -reliability

To check the agreement of two or more reviewers for a submission, a distinction can be made between agreement and interrater reliability (Kottner & Streiner, 2011). "*Agreement* points to the question, whether diagnoses, scores or judgements are identical or similar or the degree to which they differ" (Kottner & Streiner, 2011, p. 701). Reliability is concerned with "how well submissions can be distinguished from each other, in spite of measurement errors. Measurement errors are related to the variability between submissions" (de Vet et al., 2006, p. 1034). Both agreement and reliability coefficients fulfil different peer review requirements (LeBreton & Senter, 2008). If, on the one hand, one wants to know if reviewers always use the same category for a manuscript, the agreement coefficient (IRA) is the best method. If, on the other hand, manuscripts are to be categorized according to contributions with high or low ratings, then inter-rater reliability (IRR) is the method of choice. Since IRA or IRR coefficients have their specific strength and weaknesses, a set of coefficients is used (e.g., Vach & Gerke, 2023).

In addition to non-chance-corrected indices, Cohen's kappa, which indicates the degree of chance-correct agreement between two experts, has been used as an agreement coefficient (Cicchetti & Feinstein, 1990; Zhou et al., 2021). In addition to its power to correct for chance agreement, it also has limitations. Feinstein and Cicchetti (1990) describe paradoxes such as the kappa might be low even though the percentage agreement is high. Gwet's chance-corrected AC_1 was calculated in addition to Cohen's kappa, because AC_1 is paradox-resistant (Gwet, 2014, p. 104). Kappa can be very small despite high absolute agreement, especially if the frequency in a category is high. The categorizations in the two criteria "significance of findings" (e.g., landmark, fundamental) and "strength of evidence" can be ranked, which is taken into account in the weighted kappa in the form of weights according to the degree of disagreement. "Weights are usually assigned with diminishing credit, with higher values assigned for more similar pairs of experts' ratings. This is in contrast to Cohen's kappa for agreement, which only assigns credit for pairs of expert ratings that are identical" (Zhou et al., 2021, p. 4). For the weighted kappa the quadratic matrix with Cicchetti-Allison weights as entries is used. The agreement between the ratings of the first and second reviewers was determined in each case. As a submission was often rated by more than two reviewers at *eLife*, Fleiss kappa was also calculated in these cases (Zhou et al., 2021, p. 4). The singular kappa was calculated for each classification in a category, e.g., "landmark", to examine whether a submission was classified in that category or not. All possible reviewer pairs of a submission and their assessments were formed (e.g., reviewers 1 and 2, reviewers 1 and 3). A cross-table was then created for the permuted pairs to calculate the interrater agreement, with one reviewer of the pair in the row and the other reviewer of the pair in the column. The results of these analyses can be found in *Supplementary Material* (Tables S1 and S2). LeBreton and Senter (2008, p. 836) formulated revised standards for interpreting IRA estimates in a table. According to this table an IRA

of less than 0.30 indicates "lack of agreement", an IRA between 0.31 and 0.50 indicates "weak agreement", an IRA between 0.50 and 0.70 "moderate agreement", and an IRA above 0.70 or above 0.90 indicates "strong agreement" or "very strong agreement".

Intraclass correlation coefficients (ICC) have been used to determine the interrater reliability of continuous scales (Liljequist et al., 2019; Mutz et al., 2012; Shrout & Fleiss, 1979). The between-submission variance of the ratings $\left(\sigma_b^2\right)$ is related to the total variance of the ratings $\left(\sigma_{\text{tot}}^2\right)$, where the total variance consists of the between-submission variance $\left(\sigma_b^2\right)$ and the error variance $\left(\sigma_\varepsilon^2\right)$, $\text{ICC}(1,1) = \sigma_b^2 / \left(\sigma_b^2 + \sigma_\varepsilon^2\right)$ (Shrout & Fleiss, 1979, p. 423). The ICC(1, 1) estimates the single-rater reliability. "The values of 0.01 might be considered a `small` effect, a value of 0.10 might be considered a `medium` effect, and a value of 0.25 might be considered a `large` effect." (LeBreton & Senter, 2008, p. 838). To determine the composite reliability of the mean of k raters per submission, the Spearman–Brown equation was applied (k*ICC(1, 1)/(1 + (k − 1)*ICC(1, 1)) (Marsh & Bazeley, 1999, p. 13; Mutz et al., 2012, p. 3). However, the rating scales used are not continuous scales, but ordinally scaled variables. For this purpose, an intraclass coefficient was developed based on a mixed effects multinomial logistic regression model (Hedeker, 2003, p. 1439; Raykov & Marcoulides, 2015). A similar model approach was recently developed by Visscher and Yengo (2023) in the field of peer-review research.

Software

The data analyses were performed using two software packages. The SAS/STAT software (SAS Institute Inc., 2016) was used for descriptive statistics, interrater reliability and nonlinear mixed models, especially the procedures PROC FREQ/AGREE and PROC NLMIXED. The R software (R Core Team, 2024) was used to analyze interrater reliabilities (Fleiss kappa), especially the irr, icr, irrCAC, and the psych packages (Gamer et al., 2019; Gwet & Ph, 2019; Revelle, 2024; Staudt & L'Ecuyer, 2023).

Results

Comparison of the old and new publishing models

Even though the old and new publishing models are fundamentally different, it is useful to compare the two models, particularly in terms of the characteristics that have remained the same. Table 1 shows the summary statistics for some common characteristics of the two models: It distinguishes between the characteristics of the documents, the submission process, and the peer review process. To increase comparability, submissions from both models that were submitted between February and October 2019 or 2023 were included in the analysis. The results in the table show that the magnitude of the differences between the old and new publishing model is mixed.

There are no significant differences between the two models in terms of word count, the document type, and the subject area or field ("Biology & Biochemistry", "Clinical medicine"). The duration of the peer review process decreases on average from 39.8 days in the old publishing model (group 1) to 31.7 days in the new publishing model (group 3). However, the median shows an increase from 8 (group 1) to 12 days (group 3).

The results for the submission process are shown in Table 1. While the proportion of initial rejections ("editorial triage") in the new publishing model is higher (75.4%) than in

Table 1 Summary statistics

Variable	Overall (Feb-Oct)				2019 (Feb–Oct)	2023 (Feb–Oct)	
					Old model (Group 1)	Old model (Group 2)	New model (Group 3)
	M	Mdn	SD	Range	M	M	M
I. Documents							
No. of documents	10,802				6,592 (61.0%)	364 (3.4%)	3,846 (35.6%)
No. of appeals	0.04	0	0.20	0–1	0.06	0.02	0.02
No. of words in title	13.9	14	4.1	3–36	13.6	14.6	14.4
Number of peer-reviewed manuscripts	2,969				2,007 (67.6%)	87 (2.9%)	875 (29.5%)
Number of reviews per manuscript	2.73		3	1–6	2.81	2.62	2.56
Review duration [days/manuscript]	36.7	10	59.3	0–889	39.8 [Mdn 8]	31.6 [Mdn 14]	31.7 [Mdn 12]
Document type							
Research article	88.6%				87.9%	92.4%	89.4%
Short reports	5.7%				6.1%	2.8%	5.2%
Tools & resources	4.4%				4.6%	2.8%	4.2%
Research advances	1.3%				1.4%	2.0%	1.2%
Subject area							
Biology & Biochemistry	81.7%				82.7%	77.2%	80.2%
Clinical Medicine	18.4%				17.3%	22.8%	19.8%
II. Submission process							
No. of senior editors	1.0	1	0.07	1–2	1.0	1.0	1.0
No. of rev. editors	1.0	1	0.10	1–3	1.0	1.0	1.0
Initial rejection (“editorial triage”)	71.1%				68.9%	65.4%	75.4%
Final rejection	12.5%				12.8%	6.6%	-
Total rejection	79.1%				81.7%	72.0%	75.4%

Table 1 (continued)

Variable	Overall (Feb-Oct)				2019 (Feb–Oct)	2023 (Feb–Oct)	
					Old model (Group 1)	Old model (Group 2)	New model (Group 3)
	M	Mdn	SD	Range	M	M	M
Final acceptance*	20.1%				18.1%	9.9%	24.6%
III. Reviews							
No. of reviews	2.7	3	0.5	1–6	2.8	2.6	2.6
Evaluation criteria							
Significance	2.4	3	1.2	0–5	–	–	2.4
Strength	3.4	3	1.3	0–6	–	–	3.4
Scale	5.8	6	2.2	0–11	–	–	5.8

Notes. Row percents are in round brackets. M=mean, Mdn=median, SD=standard deviation

*Some manuscripts were still under consideration without a final decision at the time of data extraction

the old publishing model with 68.9% (group 1) and 65.4% (group 2), slightly more manuscripts are accepted and published in the new publishing model (24.6%) than in the old publishing model (group 1: 18.1%, group 2: 9.9%). Since all papers that pass the triage in the new model are accepted and later published by *eLife*, it seems necessary to be more rigorous in the triage of the new than in the old model.

As results published by *eLife* itself show (Ratan et al., 2023a), it is important to ask whether the introduction of the new publishing model has an effect on the distribution of subject areas among the submissions and the countries of origin of submissions (compared to the old publishing model). "In order of ranking, the highest number of submissions has been to the subject area Neuroscience, followed by Cell Biology, Genetics and Genomics, Microbiology and Infectious Diseases, and Medicine. This is similar to prior years." (Ratan et al., 2023a, p. 3). The results of the subject area comparison are shown in Fig. 1: the differences are rather small in percentage terms.

Reviewer's ratings

Interrater agreement

One measure of the quality of a rating is the extent to which reviewers rate a submission equally, i.e. using the same category on a rating scale. Tables 2 and 3 provide a $k \times k$ cross-tabulation, combining the first reviewer's rating with the second reviewer for one of the two criteria. Due to space limitations, the cross-tabulation of all possible combinations

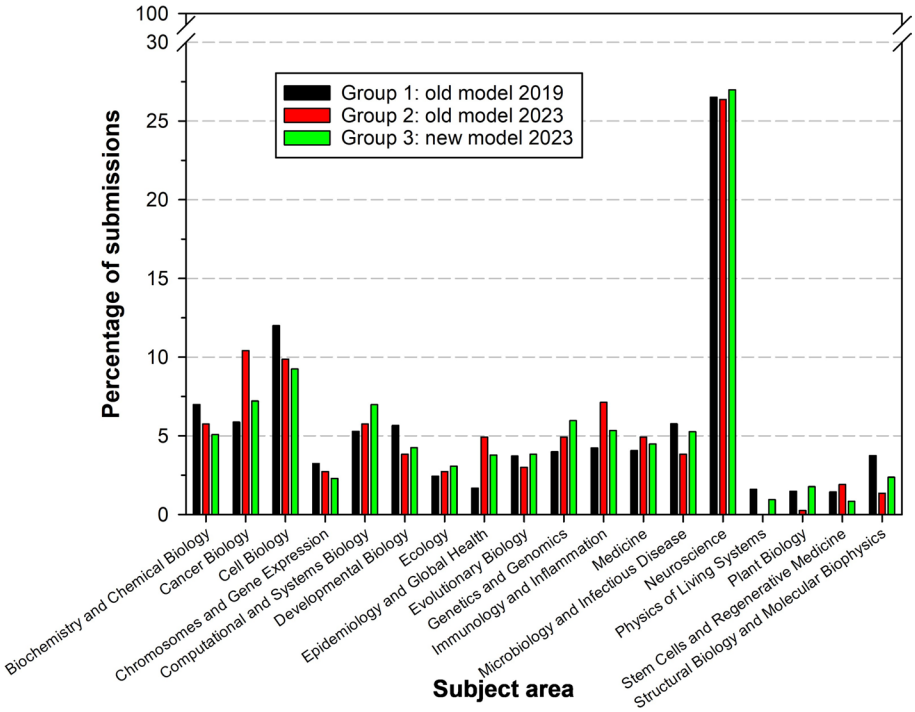

Fig. 1 Percentage of submissions in subject areas separated by groups

Table 2 Interrater agreement matrix for the first two reviewers of a manuscript regarding the reviewing scale "significance of findings" ($N = 875$ manuscripts, group 3 "new publishing model 2023")

Reviewer 1	Frequency (cell-χ^2) **Kappa**	Reviewer 2						
		6-point rating scale "significance of findings"						
		Prefer not to answer	Useful	Valuable	Important	Fundamental	Landmark	Col%
	Prefer not to answer	5 (0.40) **0.02**	7 (0.10)	15 (0.28)	23 (0.54)	4 (1.16)	2 (3.31)	6.6%
	Useful	7 (0.05)	27 (7.76*) **0.11**	34 (0.02)	35 (0.58)	8 (2.41)	2 (0.53)	13.4%
	Valuable	18 (0.16)	30 (0.51)	94 (5.0*) **0.10**	77 (0.86)	23 (1.47)	1 (0.97)	28.7%
	Important	18 (0.61)	41 (0.38)	96 (0.07)	120 (0.43) **.04**	45 (0.90)	1 (1.71)	37.9%
	Fundamental	7 (0.02)	13 (0.06)	20 (3.57)	35 (0.00)	21 (6.64*) **0.10**	3 (3.59)	11.7%
	Landmark	2 (1.18)	1 (0.48)	1 (2.54)	8 (1.91)	2 (0.05)	0 (0.15) **−0.01**	1.7%
	Row%	6.7%	14.1%	30.7%	35.2%	12.2%	1.11%	100%

Notes The number of manuscripts, the cell-χ^2 in brackets, and single Cohen's kappa (in bold) are shown in the cells. 29 manuscripts have been excluded due to a lack of ratings from the second reviewer

$^*p < .05$ χ^2 (df = 1) = 3.84; χ^2_{tot} (df = 25) = 50.4 $^*p < .05$

of reviewers of a manuscript (e.g., the first and third reviewer, or the second and third reviewer) has been omitted. In the case of perfect agreement, only the diagonal would be filled in. For the results in these tables, all submissions from 2023 for which reviews were available in the new publishing model were included ($N = 875$).

For the first criterion "significance of findings", the diagonal is most frequently occupied by individual categories such as "valuable" or "important"" but there are strong deviations from the diagonal. For example, there are 8 submissions where the first reviewer gives the rating "useful", while the second reviewer gives a significantly better rating of "fundamental". The overall low level of agreement between the reviewers is reflected in a low observed percentage agreement coefficient of 0.32 and a single Cohen's kappa corrected for chance agreement of 0.08 (Table 4). The categories "useful", "valuable" and "fundamental" still have the highest category-specific kappa coefficients of greater than or equal to 0.10 with a statistically significant cell χ^2.

The second criterion "strength of support" shows a similarly low level of agreement. Again, the observed percentage agreement is very low at 0.28 and a single Cohen's kappa of 0.08. For the categories "inadequate", "incomplete", "compelling" and "exceptional", however, there is a statistically significant cell χ^2 with a category-specific Cohen's kappa of well over 0.10 in some cases. According to LeBreton's evaluation criteria (LeBreton

Table 3 Interrater-agreement matrix for two reviewers regarding the reviewing scale "strength of support" ($N = 875$ manuscripts, group 3 "new publishing model 2023")

Reviewer 1	Frequency (cell-2) **Kappa**	Reviewer 2							
		7-point rating scale "strength of evidence"							
		Prefer not to answer	Inadequate	Incomplete	Solid	Convincing	Compelling	Exceptional	Col%
	Prefer not to answer	1 (2.28) **0.05**	1 (0.34)	2 (0.55)	5 (0.73)	3 (0.11)	1 (0.84)	1 (1.23)	1.7%
	Inadequate	1 (0.23)	7 (22.28*) **0.16**	12 (1.55)	7 (0.28)	7 (0.48)	0 (6.07*)	1 (0.02)	4.1%
	Incomplete	2 (0.38)	7 (0.00)	60 (8.03*) **0.12**	39 (0.29)	38 (1.15)	26 (0.58)	2 (1.24)	20.5%
	Solid	5 (0.43)	8 (0.03)	51 (0.00)	61 (1.73) **0.06**	53 (0.06)	31 (0.86)	2 (2.0)	24.9%
	Convincing	3 (0.34)	8 (0.24)	45 (2.45)	61 (0.16)	67 (0.48) **0.04**	48 (1.15)	5 (0.13)	27.9%
	Compelling	2 (0.16)	3 (1.52)	28 (1.76)	29 (1.60)	44 (0.65)	38 (5.50*) **0.10**	6 (1.40)	17.7
	Exceptional	1 (0.63)	0 (1.04)	5 (0.24)	5 (0.29)	8 (0.23)	3 (0.51)	4 (17.47*) **0.14**	3.1%
	Row%	1.8%	4.0%	24.0%	24.4%	26.0%	17.4%	2.5%	100%

Notes The number of manuscripts, the cell-χ^2 in brackets, and single Cohen's kappa (in bold) are shown in the cells. 29 manuscripts have been excluded due to a lack of ratings from the second reviewer

*$p < .05$ χ^2 (1) = 3.84; χ^2_{tot} (df = 36) = 91.7 *$p < .05$

Table 4 Interrater agreement coefficients of reviewer pairs [95% confidence intervals are in brackets] for group 3 "new publishing model 2023" ($N = 875$ manuscripts)

Scale	Observed agreement	Single kappa	Fleiss kappa (3 raters)	Weighted observed agreement	Weighted kappa	Chance-corrected AC_1
Significance of findings	0.32	0.08 [0.04; 0.12]	0.03* [0.00; 0.06]	0.78	0.09 [0.05; 0.14]	0.20 [0.16; 0.23]
Strength of evidence	0.28	0.08 [0.04; 0.12]	0.04* [0.01; 0.07]	0.80	0.12 [0.10; 0.19]	0.17 [0.14; 0.21]

Cicchetti–Allison-weights $(1 - |C - C_{ij}|/(C - C_{c1}))$, where c is the number of categories

*$p < .05$

& Senter, 2008), the interrater agreement coefficients for 'significance of findings' and 'strength of support' tend to indicate a lack of agreement (below 0.30).

While the single Cohen's kappa only takes into account the absolute agreement in the diagonals, the weighted kappa also includes cells of the matrix that are not on the diagonal (Table 4). In the Cicchetti–Allison weighting scheme, cells that are close to the diagonal are given a higher weight than cells that are further away from the diagonal. However, the weighting only slightly increases Cohen's kappa, for example from 0.08 to 0.12 for "strength of support". Gwet's AC1 "…represents the conditional probability that 2 raters agree given that all subjects susceptible to cause an agreement by pure chance have been removed" (Gwet, 2014, p. 68). At 0.20 and 0.17, these probabilities are slightly higher than the weighted kappa.

When all reviewers of a submission are included in the analyses (many submissions were accessed by more than two reviewers), the interrater agreement for both criteria is even lower than for the analyses with the first two reviewers of a submission (see *Supplementary Material*, Tables S1, S2).

Interrater reliability

The categories used to classify a submission can be ranked or even considered to be continuous so that interrater reliability coefficients can be calculated (Table 5). A distinction is made between the ICC for single ratings, which indicates the reliability with respect to a rating of a single reviewer, and the ICC for mean ratings, which indicates the reliability with respect to the mean of the reviewers' ratings for a submission. On average, 2.56 reviewers assessed a submission. In addition, the ICC was calculated for a composite indicator consisting of the sum of the two criteria. As these two scales have a positive correlation of 0.66 (Spearman rank correlation), a composite indicator (combining both scales) seems to be better able to discriminate between high and low quality submissions than the individual scales.

The results in Table 5 show that the ICCs for single ratings of 0.09 (continuous case) and 0.11 (ordinal case) for the criterion "significance of findings" and 0.18/0.18 for the criterion "strength of support" are comparably low as for the weighted kappa (Table 4), regardless of the scale level. The ICCs for the mean ratings are significantly higher for both "significance of findings" (0.20/0.24) and "strength of support" (0.35/0.37). The composite indicator shows no improvement in the ICCs compared to the individual criteria. According to LeBreton's evaluation criteria (LeBreton & Senter, 2008), the interrater reliability

Table 5 Single and mean interrater reliability in terms of intra-class correlation (ICC) [95% confidence intervals in brackets] for group 3 "new publishing model 2023" ($N = 875$ manuscripts)

Scale	ICC for continuous ratings		ICC for ordinal ratings	
	Single	Mean ($k = 2.56$ ratings)	Single	Mean ($k = 2.56$ ratings)
Significance of findings	0.09 [0.04; 0.14]	0.20 [0.11; 0.30]	0.11 [0.06; 0.16]	0.24 [0.14; 0.34]
Strength of support	0.18 [0.13; 0.22]	0.35 [0.27; 0.43]	0.18 [0.13; 0.23]	0.37 [0.29; 0.45]
Composite indicator	0.15 [0.10; 0.20]	0.31 [0.23; 0.40]	–	–

coefficients for ”significance of findings” and ”strength of support” indicate a medium reliability (above 0.10).

The ICC for single ordinal ratings was also calculated separately by subject area for both criteria (Fig. 2). For the criterion “significance of findings”, the ICC coefficients are above 0.20, especially for the areas “Biochemistry and Chemical Biology”, “Developmental Biology”, “Ecology”, and “Medicine”. For “Neuroscience”, which includes a large number of manuscripts, the ICC is below average.

For the criterion “strength of evidence”, the subject areas “Evolutionary Biology”, “Plant Biology”, and “Structural Biology and Molecular Biophysics” in particular have an ICC coefficient above 0.30 (Fig. 3). Here too, the ICC for “Neuroscience” is below average with a large number of manuscripts.

Besides the subject area, the ICC for single ratings was also calculated separately for each month of submission (see Figs. 4 and 5). For the criterion “significance of findings”, the ICC of 0.27 deviates significantly from the average in the months July to September. For the criterion “strength of support”, an above-average value was found not only in the months July to September, but also in February.

Determinants of reviewer's rating

In the interest of fairness, factors that are not related to the peer review process (e.g., scientific area) should not influence reviewers` ratings. A regression analysis approach was used to identify factors that influence the ratings for all submissions in 2023 for

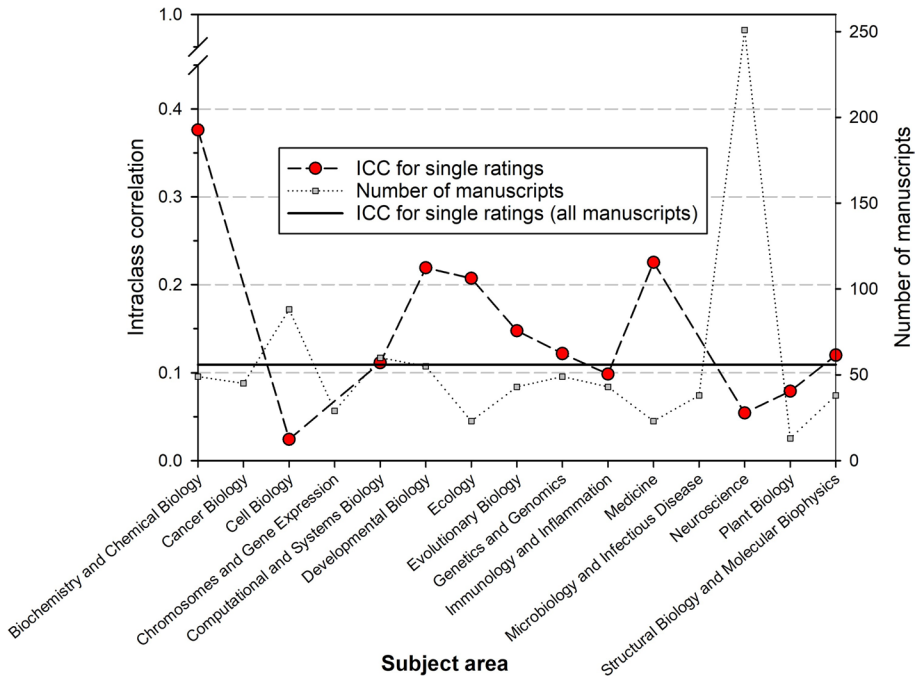

Fig. 2 Intraclass correlation for single ordinal ratings of “significance of findings” separated for areas for group 3 “new publishing model 2023”

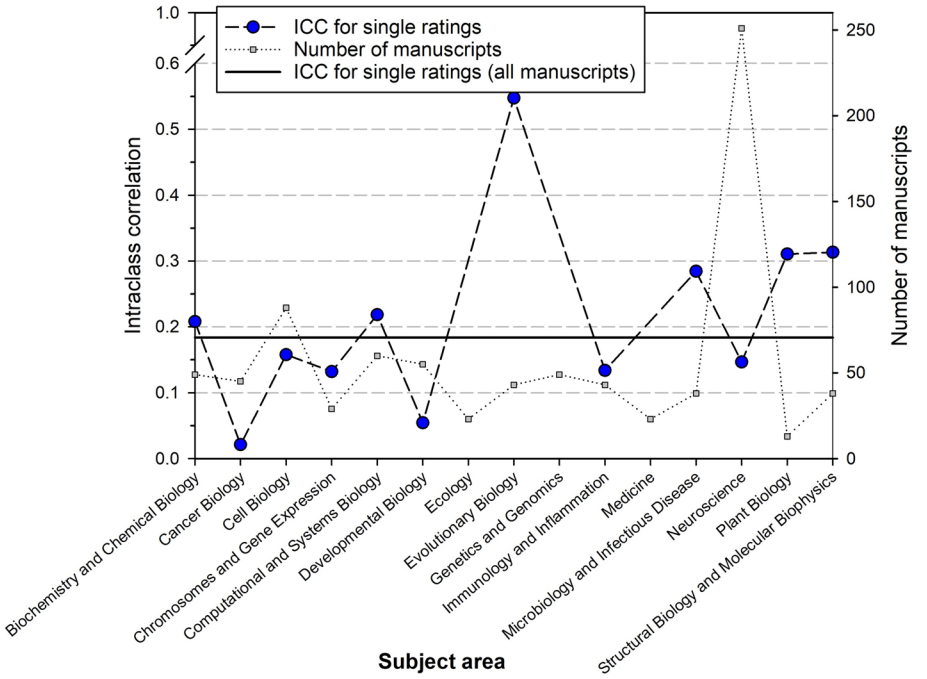

Fig. 3 Intraclass correlation for single ordinal ratings of strength of support" separated for subject areas

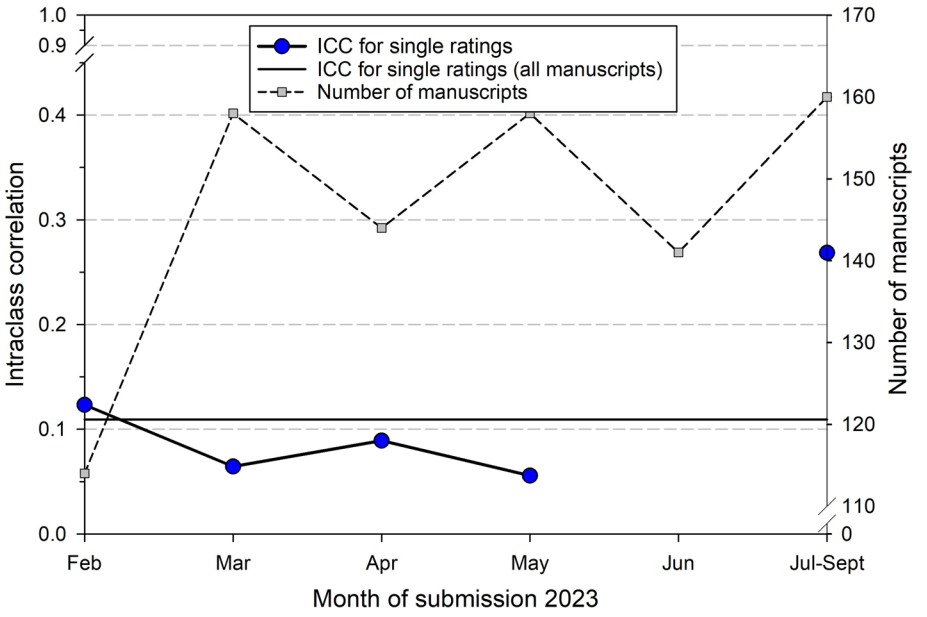

Fig. 4 Intraclass correlation for single ordinal ratings of "significance of findings" separated for each month of submission

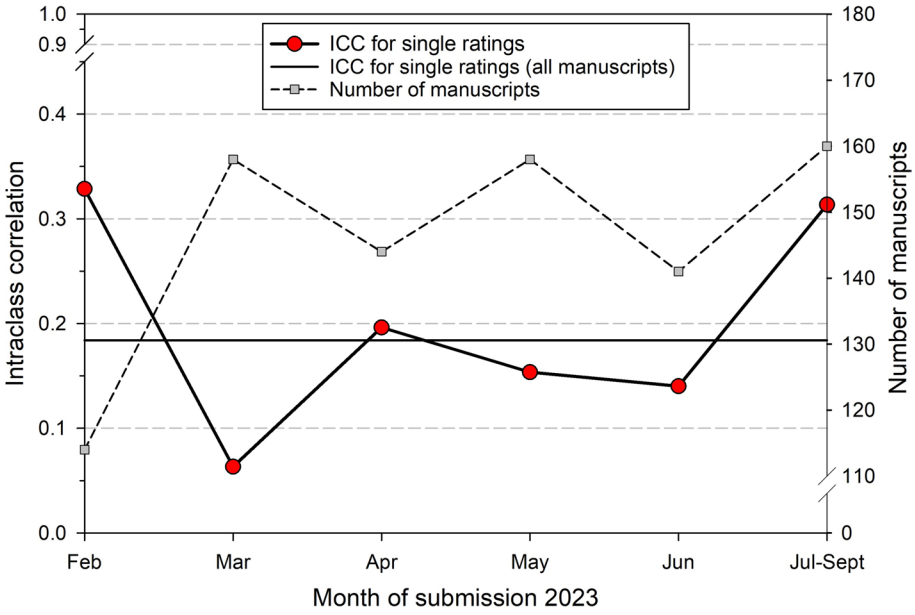

Fig. 5 Intraclass correlation for single ordinal ratings of "strength of support" separated for each month of submission

which reviews were available. The same model that was used to calculate the intraclass correlation for ordinal data (Table 5) was also used for analysing fairness (multilevel ordinal logistic regression). The following six characteristics of the submission were considered: number of revisions for a manuscript, scientific fields ("neuroscience", "cell biology"), number of reviews and whether the review duration is above average (= 1) or not. When using the Bayes information criterion (BIC) to statistically compare models with and without covariates, the models with covariates is worse than the model without covariates for both criteria ($BIC_{signif} = 6{,}584.6$ vs. $BIC_{signif} = 6{,}605.4$; $BIC_{strength} = 7{,}204.9$ vs. $BIC_{strength} = 7{,}223.4$). The smaller the BIC, the better the fit in relative terms (Schwarz, 1978). The fairness of open peer review was not or only slightly compromised.

As the rating scales consist of ranked categories (e.g., "useful", "valuable", …), the threshold values (τ_1, τ_2, …) for the transitions between the categories are highly differentiated for both criteria (−2.84 to 4.72, 4.76 to 3.57). The variance of the random variable σ_u^2 reflects the variability between the manuscripts. If it were zero, reviewers would on average not agree in their judgments for a manuscript. Overall, only 10% and 7% of the variance of the random intercept (R^2) of the two criteria are explained by the six mean-centred factors. For the first criterion "significance of findings", "number of revisions" ($\beta_2 = -0.20$), "Cell Biology" ($\beta_4 = -0.35$) and "review duration above mean" ($\beta_6 = 0.21$) are statistically significant. Since statistical procedure models the probabilities of the criteria values that are ranked lower, submissions from the field "Cell Biology" that have undergone an above-average number of reviews perform better than submissions that do not have these characteristics. However, submissions with an above-average review duration scored poorly. With regard to the "strength of support"

Table 6 Parameter estimates of the multilevel ordinal logistic regression model for the pair of review criteria

Effect	Param	Significance of findings			Strength of support		
		Estimate	SE	*t*-value	Estimate	SE	t-value
Thresholds							
Category 0	τ_1	−2.84	0.17	−16.9*	−4.76	0.23	−20.49*
Category 1	τ_2	−1.46	0.15	−9.72*	−3.54	0.19	−18.70*
Category 2	τ_3	0.03	0.14	0.22	−1.49	0.16	−9.13*
Category 3	τ_4	2.07	0.16	13.27*	−0.23	0.16	−1.42
Category 4	τ_5	4.72	0.25	18.70*	1.22	0.16	7.50*
Category 5	τ_6	–	–	–	3.57	0.21	17.28*
Fixed effects (submissions)							
Research article	β_1	−0.06	0.14	−0.42	0.30	0.15	1.96
No. of revisions	β_2	−0.20	0.10	−2.08*	−0.14	0.11	−1.27
Neuroscience (=1)	β_3	−0.04	0.10	−0.40	0.03	0.11	0.30
Cell biology (=1)	β_4	−0.35	0.15	−2.28*	−0.15	0.17	−0.91
No. of reviews	β_5	0.05	0.09	0.53	−0.05	0.10	−0.51
Review duration above mean (=1)	β_6	0.21	0.10	2.15*	0.33	0.11	3.11*
Random effects							
Random intercept	σ_u^2	0.37	0.11	3.38*	0.69	0.13	5.37*
ICC_{single}	ICC(1)	0.10 [0.05; 0.15]			0.17 [0.12; 0.23]		
$ICC_{average}$	ICC(k)	0.22 [0.12; 0.32]			0.35 [0.27; 0.43]		
Explained variance	R^2	0.10			0.07		
BIC		6,605.4			7,223.4		

The statistical procedure is modelling the probabilities of criteria levels having lower ordered values in the response profile (e.g., "prefer not to answer", "useful", "valuable", "important", …). The 95% confidence intervals are in brackets: SE=standard error

*$p < .05$ ($df = 874$)

criterion, the rating drops if the review duration is above average. The consideration of covariates does not have a significant influence on the intraclass correlations, if one compares the model with covariates (Table 6) and without covariates (Table 5).

Summary and discussion

eLife's new publishing model, which was introduced at the end of January 2023, breaks with the previous traditional model: all manuscripts submitted as preprints are reviewed and published if they are deemed worthy of review in an initial "editorial triage" (Abbott, 2023; Eisen et al., 2020, 2022; Graham, 2022; Urban et al., 2022). The new model is "publish, then review" (Eisen et al., 2020, p. 1). The gatekeeping function of peer review is abandoned in favor of its improving function, i.e. a "consultative approach to peer review" (King, 2017; Schekman, 2017, p. 1). The introduction of the new publishing model can be interpreted as a necessary improvement of peer review, since the results of research on journal peer review have repeatedly shown a low

interrater reliability of reviewers' ratings, as a meta-analysis by Bornmann et al. (2010) shows. Reasons for the lack of interrater reliability could lie in the problem of noise as "unwanted variability in judgments", which Kahneman (2022, p. 24) refers to in an interview regarding his book "Noise: A flaw in human judgment" (Kahneman et al., 2021). Like journal reviewers in a journal peer review processes for single submissions, doctors and judges come to completely different conclusions about the same case. Given this evidence for expert judgement in peer review and other decision-making processes, the use of expert judgement for final decisions on whether or not to publish a manuscript seems questionable.

Since the new publishing model at *eLife* encourages the readers (but not the reviewers) to decide whether or not to read (use) a publication, peer review in the form of (public) reviewers' ratings and comments continues to have the "signaling" function (e.g., Franck, 2002). This function of peer review reduces "information asymmetries (e.g., the editor knows how the referees' rate an article but the readers of the paper do not) by highlighting extraordinary quality or relevance and … [fosters] efficient allocation of researcher's scarcest resource in the age of information—namely attention" (Mutz et al., 2017, p. 2139f).

The results of this study can be summarized as follows in accordance with the four questions in the introduction section:

1. *Descriptive comparison of publishing models*: What are the differences between the old and new publishing models in terms of submission characteristics? There are only small differences between the submissions of the two publishing models in the available variables, as the number of words, the document type, and the subject area (e.g., "Biology & Biochemistry", "Clinical Medicine"). *eLife* receives manuscripts from different subject areas with comparable frequency in the old and new publishing models.
2. *Interrater reliability of peer review*: How well do reviewers agree in their assessments of manuscripts submitted under the new publishing model with respect to the review criteria "significance of finding" and "strength of support"? The results in this study point to a low level of agreement. The Cohen's kappa corrected for chance agreement is 0.08 [0.04; 0.12] for both criteria. According to LeBreton's criteria (LeBreton & Senter, 2008), there is a lack of absolute agreement between the experts. The weighted kappa is slightly higher. The interrater reliability (ICC for single ratings), which is based on continuous scales, is also low at 0.09 [0.04; 0.14] for "significance of findings" and 0.18 [0.13; 0.22] for "strength of support". According to LeBreton's evaluation criteria (LeBreton & Senter, 2008), the interrater reliability coefficients for "strength of support" tend to indicate a medium reliability (above 0.10). If the mean ratings of all reviewers of a submission are used instead of the individual reviewer ratings per submission, the interrater reliability increases to 0.20/0.35. However, the interrater agreement remains at a low level.
3. *Fairness of peer review*: Are there any factors that influence peer review ratings that are part of the formal peer review process? A regression analysis was used to identify factors that influence the peer review ratings for all submissions in 2023 for which reviews were available. Since the model with covariates is worse than the model without covariates, the fairness of peer review seems to be not or only slightly compromised. Only 10% and 7% of the random intercept variance (R^2) for the two criteria is explained by the six factors.

What are the limitations of this study? Since only the data from one year was collected in both the old and the new publishing models, the generalizability of the results is limited. The old and the new models are only comparable to a limited extent. For example, the old publishing model lacks quantitative judgments in the peer review process. This study focuses only on the quantitative judgements; the qualitative judgements of the reviewers (their texts) have not been analyzed. This is mainly due to the fact that they rather contain arguments for improving the paper and less evaluations. Following previous studies (e.g., Prabowo & Thelwall, 2009), these arguments may be the subject of a sentiment analysis in the future.

Based on our empirical results, the following recommendations can be derived for the peer review process at *eLife*:

- *Number of reviewers*: The more reviewers are involved in the peer review process, the more reliable the assessments will be. Reviewer's ratings might not be perfectly reliable and may fluctuate randomly around a "true" estimate of assessment. The reliability of the mean ratings increases, the more reviewers are included (Marsh & Bazeley, 1999, p. 13). In order to have a reliable "signaling" function of peer review, it would be worth considering at *eLife* to increase the number of reviewers per paper (e.g., at least 3–4). Currently, an average of 2.7 referees assess a manuscript. Even if the rating scales for the assessments are ordinal, the journal could calculate means (or medians) for the published papers to provide "quality signals".
- *Signaling*: In the new publishing model without gate keeping the *eLife* assessments (and information from the peer review process) become more important to reflect the quality differences between the published papers. However, Ratan et al. (2023a) found that 50% of readers do not use the information provided by the reviewers. We therefore recommend that *eLife* emphasizes stronger the importance of this information to the reader. We also recommend in this regard the use of "signals" by the journal that are derived from the peer review process. Specific labels can alert readers to exceptional papers (Mutz et al., 2017). For example, the journal *Angewandte Chemie International Edition* highlights exceptional papers as "very important papers" (VIPs) or "hot papers". Mutz et al. (2017) examined the citation impact of communications with and without VIP labels. A statistical causal analysis using propensity score matching showed that papers with a VIP label received around 20 more citations than a comparable group of papers without VIP label.
- *Editorial triage*: The decision on whether or not to publish in *eLife* is made by a group of editors: in the old publishing model partly with the support of peer review, and in the new model completely without peer review. In the new model, the importance of editorial triage has increased: it is the only step that decides about the publication fate of submissions at *eLife*, and this step may be prone to biases. In the context of Open Science in Psychology, Sharpe (2024) speaks of an editor bias: "Editor bias is when editors fail to be fair and impartial in their handling of articles" (Sharpe, 2024, p. 883). As our results show that more manuscripts were rejected by the editors in the new publishing model than in the old one, questions arise about the quality of the editors' decisions. In accordance with Sharpe (2024), we recommend that the triage process at *eLife* be made more transparent to the public to avoid speculation about unfairness: authors should know the criteria used to decide whether or not to send a manuscript for peer review. If *eLife* decides to completely abandon the gatekeeping function of peer review (i.e. editorial decisions after peer review), all submitted manuscripts (i.e. including those rejected in the triage) should be made public including the reasons why manuscripts

were or were not sent out for peer review. This decision for full transparency would lead to an open publication and peer review process to the journal which may also discourage the submission of manuscripts with poor quality in the future. Authors of these manuscripts would know the risk that others would become aware of their poor quality research and the journal's negative decision.

In order to prevent possible unfairness in the editorial triage two recommendations by Sharpe (2024, p. 883) could be considered in the decision process: "masking author identity" and "increasing editor diversity". Both actions may not only increase the fairness of the process, they would also publicly demonstrate that *elife* is aware of possible problems with unfairness in the process and tries to tackle that.

The validity and fairness of the editorial triage and the expert ratings as well as the usefulness and quality of the reviewers' text contributions were not analyzed in this study. Questions of validity can be answered by correlating editorial decisions and expert ratings with external quality criteria such as citations. As this requires a citation window of at least 3 years to arrive at reliable citation counts, the validity of the peer review process could not be analyzed in this study. It would also be very helpful for the analysis of bias and unfairness of the editorial triage und the expert ratings if citation data were available.

Research on bias distinguishes between potential and real bias. In most bias studies, there is only evidence that manuscripts from one social group (e.g., men) receive better reviews and are more likely to be published than manuscripts from another social group (e.g., women). This is known as potential bias. In order to be able to speak of real bias, one would first have to examine the fate of manuscripts that were rejected but published elsewhere. A proxy measure of quality (e.g., citation frequencies) would then be determined for both groups of manuscripts. If, for example, the mean reviewer ratings for two groups of manuscripts were different, but the mean citation frequencies for both manuscript groups were not different, there would be a real bias. The usefulness of reviewers` input could be checked, for example, by asking the corresponding authors in an online survey.

We plan to address the issues of validity, fairness as well as the usefulness and quality of the reviewers' text contributions in follow-up studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-025-05422-y.

Author contributions All authors contributed to the conceptualization, design and method. Data collection and formal analysis and investigation, writing the original draft, preparation and visualization were performed by Rüdiger Mutz. Review editing, funding acquisition, resources and supervision were performed by Hans-Dieter Daniel and Lutz Bornmann.

Funding Open access funding provided by University of Zurich.

Data availability The data is available on request from *elife* editorial board.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbott, A. (2023). Strife at *eLife*: Inside a journal's quest to upend science publishing. *Nature, 615*(7954), 780–781. https://doi.org/10.1038/d41586-023-00831-6

Aczel, B., Barwich, A. S., Diekman, A. B., Fishbach, A., Goldstone, R. L., Gomez, P., Gundersen, O. E., von Hippel, P. T., Holcombe, A. O., Lewandowsky, S., Nozari, N., Pestilli, F., & Ioannidis, J. P. A. (2025). The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences of the United States of America, 122(5): e2401232121*. https://doi.org/10.1073/pnas.2401232121

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics, 3*(1), 63–84. https://doi.org/10.1007/s11948-997-0017-3

Bornmann, L. (2015). Interrater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for Information Science and Technology, 66*(12), 2415–2426. https://doi.org/10.1002/asi.23334

Bornmann, L., & Daniel, H. D. (2008a). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie. *Angewandte Chemie - International Edition, 47*(38), 7173–7178. https://doi.org/10.1002/anie.200800513

Bornmann, L., & Daniel, H. D. (2008b). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59*(11), 1841–1852. https://doi.org/10.1002/asi.20901

Bornmann, L., & Daniel, H. D. (2009). Reviewer and editor biases in journal peer review: An investigation of manuscript refereeing at Angewandte Chemie International Edition. *Research Evaluation, 18*(4), 262–272. https://doi.org/10.3152/095820209X477520

Bornmann, L., & Daniel, H. D. (2010a). *Predictive validity of editorial decisions at an open access journal: A case study on Atmospheric Chemistry and Physics.* Paper presented at the ELPUB 2010 - Publishing in the Networked World: Transforming the Nature of Communication, 14th International Conference on Electronic Publishing.

Bornmann, L., & Daniel, H. D. (2010b). Reliability of reviewers' ratings when using public peer review: A case study. *Learned Publishing, 23*(2), 124–131. https://doi.org/10.1087/20100207

Bornmann, L., & Daniel, H. D. (2010c). The validity of staff editors' initial evaluations of manuscripts: A case study of Angewandte Chemie International Edition. *Scientometrics, 85*(3), 681–687. https://doi.org/10.1007/s11192-010-0215-7

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS One 5(12): e14331.* https://doi.org/10.1371/journal.pone.0014331

Cheng, X., Wang, H., Tang, L., Jiang, W., Zhou, M., & Wang, G. (2024). Open peer review correlates with altmetrics but not with citations: Evidence from Nature Communications and *PLOS One*. *Journal of Informetrics, 18(3), 101540*. https://doi.org/10.1016/j.joi.2024.101540

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119–186.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa 2 Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-m

Daniel, H.-D. (1993). *Guardians of science: Fairness and reliability of peer review*. Wiley.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Eisen, M. B., Akhmanova, A., Behrens, T. E., Diedrichsen, J., Harper, D. M., Iordanova, M. D., Weigel, D., & Zaidi, M. (2022). Peer review without gatekeeping. *eLife 9: e64910*. https://doi.org/10.7554/eLife.83889

Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., & Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. *eLife*, December 1, 2020. https://doi.org/10.7554/eLife.64910

eLife Editorial, L., eLife Senior, E., & eLife Early Career Advisory, G. (2024). How and why eLife selects papers for peer review. *eLife 9: e64910*. https://doi.org/10.7554/eLife.100571

Else, H. (2022). eLife won't reject papers once they are under review—What researchers think. *Nature,* 03 November 2022. https://doi.org/10.1038/d41586-022-03534-6

Enders, C. K. (2025). Missing data: An update on the state of the art. *Psychological Methods, 30*(2), 322–339. https://doi.org/10.1037/met0000563

Feinstein, A. R., & Cicchetti, D. V. (1990). High aggreement but low Kappa. 1 The problems of 2 paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-l

Franck, G. (2002). The scientific economy of attention: A novel approach to the collective rationality of science. *Scientometrics, 55*(1), 3–26. https://doi.org/10.1023/A:1016059402618

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement-R package version 0.84.1. https://CRAN.R-project.org/package=irr

Graham, F. (2022). Daily briefing: *eLife* won't reject papers under review. *Nature,* November 4, 2022. https://doi.org/10.1038/d41586-022-03600-z

Gwet K.L., & Ph. D. (2019). irrCAC: Computing chance-corrected agreement coefficients (CAC) R package version 1.0. https://CRAN.R-project.org/package=irrCAC

Gwet, K. L. (2014). *Handbook of inter-rater reliability-The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hardwicke, T. E., Schiavone, S. R., Clarke, B., & Vazire, S. (2024). An empirical appraisal of *eLife*'s assessment vocabulary. *PLoS Biology, 22*(8): e3002645. https://doi.org/10.1371/journal.pbio.3002645

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine, 22*(9), 1433–1446. https://doi.org/10.1002/sim.1522

Horbach, S. (2021). No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Research Evaluation, 30*(3), 231–239. https://doi.org/10.1093/reseval/rvaa037

Kahneman, D. (2022). Try to design an approach to making a judgment; Don't just go into it trusting your intuition. *Issues in Science and Technology, 38*(3), 23–26.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology, 64*(6), 701–702. https://doi.org/10.1016/j.jclinepi.2010.12.001

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2–17. https://doi.org/10.1002/asi.22784

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation—A discussion and demonstration of basic features. *PLoS One 14*(7): e0219854. https://doi.org/10.1371/journal.pone.0219854

Marder, E. (2014). In numbers we trust? *eLife, 3*: e02791. https://doi.org/10.7554/eLife.02791

Marsh, H. W., & Bazeley, P. (1999). Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. *Multivariate Behavioral Research, 34*(1), 1–30. https://doi.org/10.1207/s15327906mbr3401_1

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160

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): e48759. https://doi.org/10.1371/journal.pone.0048759

Mutz, R., Bornmann, L., & Daniel, H. D. (2015). Testing for the fairness and predictive validity of research funding decisions: A multilevel multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian Science Fund. *Journal of the Association for Information Science and Technology, 66*(11), 2321–2339. https://doi.org/10.1002/asi.23315

Mutz, R., Wolbring, T., & Daniel, H. D. (2017). The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology, 68*(9), 2139–2153. https://doi.org/10.1002/asi.23701

Nicolai, A. T., Schmal, S., & Schuster, C. L. (2015). Interrater reliability of the peer review process in management journals. In I. M. Welpe, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and performance* (pp. 107–119). Springer.

Patterson, M., & Schekman, R. (2018). A new twist on peer review. *eLife*, *7*: e36545. https://doi.org/10.7554/eLife.36545

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

R Core Team. (2024). R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ratan, K., Greene, S., & Rele, S. (2023a). eLife new model at six months: An ICOR analysis. September 28, 2023. https://elifesciences.org/inside-elife/7776315b/elife-new-model-at-six-months-an-icor-analysis

Ratan, K., Greene, S., & Rele, S. (2023b). eLife's new model: Initial three-month update. May 16, 2023. https://elifesciences.org/inside-elife/739ce507/elife-s-new-model-initial-three-month-update

Raykov, T., & Marcoulides, G. A. (2015). Intraclass correlation coefficients in hierarchical design studies with discrete response variables: A note on a direct interval estimation procedure. *Educational and Psychological Measurement, 75*(6), 1063–1070. https://doi.org/10.1177/0013164414564052

Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research - R package version 2.4.1. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych

Rodgers, P. (2017). Plain-language summaries of research: Writing for different readers. *eLife, 6*: e25408. https://doi.org/10.7554/eLife.25408

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research*, *6*: 588. https://doi.org/10.12688/f1000research.11369.2

Ross-Hellauer, T., Bouter, L. M., & Horbach, S. P. J. M. (2023). Open peer review urgently requires evidence: A call to action. *PLoS Biology*, *21*(10): e3002255 https://doi.org/10.1371/journal.pbio.3002255

SAS Institute Inc. (2016). *SAS for Windows*. SAS Institute Inc.

Schekman, R. (2017). Room at the top. *eLife*, October 12, 2017. https://doi.org/10.7554/eLife.31697

Schekman, R. (2019). Progress and promise. *eLife,* January 23, 2019. https://doi.org/10.7554/eLife.44799

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sharpe, D. (2024). Editor bias and transparency in psychology's open science era. *American Psychologist, 79*(7), 883–892. https://doi.org/10.1037/amp0001224

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178–182. https://doi.org/10.1258/jrsm.99.4.178

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

Staudt, A., & L'Ecuyer, P. (2023). icr: Compute Krippendorff's Alpha R package version 0.6.4. https://CRAN.R-project.org/package=icr

Sun, Z., Clark Cao, C., Ma, C., & Li, Y. (2023). The academic status of reviewers predicts their language use. *Journal of Informetrics, 17*(4), 1. https://doi.org/10.1016/j.joi.2023.101449

Urban, L., De Niz, M., Fernández-Chiappe, F., Ebrahimi, H., Han, L. K., Mehta, D., Mencia, R., Mittal, D., Ochola, E., Paz Quezada, C., Romani, F., Sinapayen, L., Tay, A., Varma, A., Yahia Mohamed Elkheir, L., & Elkheir, L. Y. M. (2022). eLife's new model and its impact on science communication. *eLife, 11*: e84816. https://doi.org/10.7554/ELIFE.84816

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa—A comparison of basic properties. *MethodsX, 10*:102212 https://doi.org/10.1016/j.mex.2023.102212

Vaggi, F., Giordan, M., Csikasz-Nagy, A., & Collings, A. M. (2016). The effects of an editor serving as one of the reviewers during the peer-review process. *F1000Research*, *5*: 683 https://doi.org/10.12688/f1000research.8452.2

Visscher, P. M., & Yengo, L. (2023). The effect of the scale of grant scoring on ranking accuracy. *F1000Research*. *11*:1197. https://doi.org/10.12688/f1000research.125400.2

Walker, R., & Rocha da Silva, P. (2015). Emerging trends in peer review-A survey. *Frontiers in Neuroscience, 9*(APR), 169. https://doi.org/10.3389/fnins.2015.00169

Waltman, L., Kaltenbrunner, W., Pinfield, S., & Woods, H. B. (2023). How to improve scientific peer review: Four schools of thought. *Learned Publishing, 36*(3), 334–347. https://doi.org/10.1002/leap.1544

Zhou, T. J., Raza, S., & Nelson, K. P. (2021). Methods of assessing categorical agreement between correlated screening tests in clinical studies. *Journal of Applied Statistiscs, 48*(10), 1861–1881. https://doi.org/10.1080/02664763.2020.1777394