Article

active learning in higher education

Active Learning in Higher Education
1–18

DOI: 10.1177/14697874241264495
journals.sagepub.com/home/alh
Sage

Engaging undergraduate students in preprint peer review

Dawn Holford
University of Bristol, UK

Janet McLean
University of Abertay, UK

Alex O Holcombe
University of Sydney, Australia

Iratxe Puebla
DataCite, Hannover, Germany

Vera Kempe
University of Abertay, UK

Abstract

Authentic assessment allows students to demonstrate knowledge and skills in real-world tasks. In research, peer review is one such task that researchers learn by doing, as they evaluate other researchers' work. This means peer review could serve as an authentic assessment that engages students' critical thinking skills in a process of active learning. In this study, we had students write peer reviews of preprints, scaffolded by a rubric. Agreement between the students and academics was reasonable, and active student involvement was high. The results suggest that use of peer review in undergraduate classes should be explored more. It likely facilitates students' ability to evaluate the quality of scientific studies, encourages active learning about the scientific process and shows potential for contributing to publicly-available assessment of scientific studies.

Keywords

authentic assessment, peer review, preprints, preprint review, rubric tool

This study introduces a tool to support training students in peer review of preprints and evaluates whether peer review aided in this way is suitable as an authentic assessment for undergraduate students (we refer to peer review in the context of scholarly publishing rather than students

Corresponding author:
Dawn Holford, School of Psychological Science, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, UK.
Email: dawn.holford@bristol.ac.uk

reviewing each other's assessments). An authentic assessment involves completing a task that serves a need or fulfils a purpose beyond the mere assessment of student learning, such as fulfilling a societal need (McArthur, 2023). Peer review training might lend itself especially well to authentic assessment in higher education by allowing students to complete such a task (Sokhanvar et al., 2021). Moreover, peer review could foster active involvement in this learning task if they perceive it to be meaningful beyond the requirements of the curriculum. We posit that to complete a peer review, students must actively apply knowledge acquired in class, while consolidating their learning and gaining insight about scientific theories and methods. Active involvement facilitates self-reflection and self-efficacy and has been shown to have beneficial effects for learning and beyond, especially in adolescents and young adults (Greene, 2013).

Introducing undergraduate students to the process of scholarly peer review can raise their awareness of current challenges in scientific publishing and endow them with evaluative skills to make informed decisions about the quality of scientific evidence (McDowell et al., 2022). These are key skills in an age of rapid and open information sharing (Lima & Nascimento, 2022). This transferable skill can be applied to scrutinising a range of sources beyond scientific reports. This is particularly important because recent efforts to speed the dissemination of scientific findings have extended the practice of publicly posting draft manuscripts on preprint servers to the biological (e.g. bioRxiv, created 2013), medical (e.g. medRxiv, created 2019) and social sciences (e.g. PsyArXiv, created 2016). In times of urgent societal challenges, like the COVID-19 pandemic, many journalists, policy makers and the lay public relied on preprints for rapidly emerging evidence. Yet roughly one-third of preprints may never appear in a journal and, hence, may not have undergone vetting by the scientific community (Anderson, 2020; Eckmann & Bandrowski, 2023). This emphasises the need for graduates to apply the skill of critical review to research reports.

Peer review is inherently an active learning process, which generations of academic reviewers have learned by doing (Munasinghe et al., 2022). For undergraduate students, gaining familiarity with the peer review process can also provide an active experience of how reviewers advance self-correction in the research community as part of the scientific method. Preprints provide a prime opportunity for active learning because, unlike scholarly manuscripts submitted to a journal, they are freely accessible and open for anyone to review. In fact, preprints can help authors receive feedback to improve their draft manuscripts (Sever et al., 2019) prior to their subsequent submission to a peer-reviewed journal; to this end, many preprint servers provide commenting features (Kirkham et al., 2020). Commenting on a draft that may still be subject to change can be motivating for students and incentivise them to provide their critique. If students can acquire reviewing skills, it would provide a much-needed service to the research community (Warne, 2016).

Reviewer training is not a new idea, but the use of peer review of preprints as a meaningful pedagogical tool that involves undergraduate students specifically in active learning about the scientific publishing process has not been widely explored. PREreview, a platform to encourage reviewing of preprints and broaden the peer reviewer base (Hindle & Saderi, 2017), targets early career researchers and supports communities of reviewers with review rubrics. The ASAPbio organisation also facilitates reviewing but is targeted at 'crowd review' communities of volunteers independent of career stage (Puebla, 2022). Moreover, studies looking at the effect of journal reviewer training on producing better quality reviews have yielded mixed results. Two meta-analyses that examined the effect of reviewer training in a small number of randomised controlled trials found no measurable effect on review quality (Bruce et al., 2016) and a moderate positive effect at a cost of increased duration (Gaudino et al., 2021); both studies concluded that sufficient evidence is lacking. These findings may, amongst other things, be due to the level of prior experience by participating reviewers and thus limit the potential gains from training. There does not seem to be an agreed standard for how to train the skill of peer reviewing, how to evaluate its efficacy and how to assess students in this skill.

In addition, previous initiatives were not explicitly embedded into a structured training programme at the undergraduate level. We know of only one exploratory study that reported on successful inclusion of peer review within an undergraduate biology curriculum for 19 students (Otto et al., 2023). The structure of undergraduate programmes can present considerable challenges for teaching peer review because accredited undergraduate degrees are required to deliver specific modules aligned with their programme structures, leaving limited opportunity to add an additional module focused exclusively on peer reviewing. Additionally, to evaluate new research, students require some subject competence, which they learn in subject-specific modules, for example, Cognitive Psychology or Language Development. To overcome these challenges, and to foster active learning about the scientific process in their discipline, we introduced peer review as an authentic assessment into a subject-based module such that preparation for the assessment constitutes training in peer-review. To do so, we started by creating a tool that can guide students through the process of preprint review writing and is flexible enough to be embedded into existing university modules as an authentic assessment.

Our approach involved undergraduate students writing a peer review of a preprint, scaffolded by a rubric that catalogued content elements of preprint reviewing (rather than providing evaluative criteria for the student's work). The scaffold made explicit the task was not just to describe the research, but to evaluate its quality. While some organisations include evaluation questions to help peer reviewers structure their critique (e.g. Foster et al., 2021; Fraser et al., 2023), these are typically brief and/or focus on one element of the evaluation. Breaking down the process and the multiple considerations that a reviewer might have, including tacit knowledge among academics, required a more detailed structure than we were able to find. Therefore, our rubric highlighted different aspects of research that need to be scrutinised, giving structure to what is otherwise a complex novel task for students. The rubric was designed to help extract and evaluate relevant information from a research report, thus scaffolding the process of writing one's own review.

The rubric, described in detail below, was provided to all the students in a final (4th)-year undergraduate module that was part of programmes in Psychology and Social Science (40–60 students) in two consecutive academic years. Students were required to fill out the rubric prior to writing a peer review on their own as an active learning exercise. Four of the authors (DH, VK, JM, AH) who qualify as expert academic reviewers in the field of psychology also wrote a peer review after completing the rubric. Specifically, the study was designed to explore the following research questions:

1. Will students be actively involved when given our authentic preprint peer review assessment? We submitted the students' work to Turnitin™, which yielded Similarity Scores (henceforth: Turnitin™ scores) indicating the percentage of text in each student's work that overlaps with sources on the internet or other students' work. Low Turnitin™ scores indicate that students neither copied from sources available online nor from each other. The assignment was also completed before generative AI was widely available. Therefore, low Turnitin™ scores mean that students generated authentic text themselves and can serve as a proxy for students being actively involved in the completion of the task (du Rocher, 2020; Murdock & Anderman, 2006).
2. Does the peer review rubric enable undergraduate students to extract relevant information from preprints? This question was answered by comparing students' rubric responses with academics' (as an indication of the quality of student evaluations) and by comparing

subjective ratings of the difficulty with which relevant information from preprints can be extracted and evaluated.

3. Does the peer review rubric guide students to write preprint reviews with content similar to academics' reviews? If so, one would expect the level of similarity between student and academic reviews (measured by linguistic similarity scores) to be associated with two rubric measures: (1) the number of rubric items mentioned in student reviews, and (2) the rate of agreement between student and academic rubric responses. Note that we did not investigate whether the rubric would improve student preprint reviews compared to reviews written without first completing the rubric. It was difficult to justify having students complete a complex task, on which they would be graded, without the rubric to provide some support. Rather, we were interested in the extent to which, with the rubric, undergraduate student work would approximate academics' reviews.

In answering these research questions we refrain from using instructor grades of the student review assessment as review quality indicators because grades constitute a subjective and hence unreliable quality measure, and lack construct validity. Instructors are likely to evaluate student reviews differently than a journal editor would, since they would make allowances for what constitutes good student work rather than a good review.

Method

Ethical approval was obtained from the University of Abertay prior to the commencement of the teaching modules. We pre-registered the rubric, the data collection procedure, and primary analyses (available from https://doi.org/10.17605/OSF.IO/X9UTN; (Holford, Kempe, Holcombe, Puebla, Patel, Smout and McLean, 2021)) prior to testing Cohort 1. At the time of pre-registration, it was not known whether subsequent cohorts could be tested, as this was contingent on the successful completion of the assessment of Cohort 1. We therefore only pre-registered testing of Cohort 1 but subsequent testing of Cohort 2 adhered to the same protocol except for some small amendments mentioned below. These changes mean that all analyses involving both cohorts and including Cohort as a fixed effect could be considered exploratory rather than pre-registered. Moreover, at the beginning of the study, we were not clear on how to measure student active involvement beyond anecdotal evidence and student comments in university module evaluations, but over the course of the project, we realised that Turnitin™ scores may serve this purpose. After receiving separate ethical approval for use of these data and for the inclusion of historical Turnitin™ scores from the cohort prior to this study, we report these scores in an exploratory analysis.

Participants

We recruited undergraduate students enrolled in a final (4th)-year undergraduate psychology module that taught topics relevant to early years education: language development, literacy, numeracy development, play and use of media. The study was conducted with two student cohorts in consecutive academic years (2021–2022 and 2022–2023). Students completed the study as a module requirement but provided informed consent to the use of their data for research (with no penalty to those who opted out). After excluding one student in Cohort 1 and two students in Cohort 2 who did not consent, our sample size was 60 in Cohort 1 and 36 in Cohort 2. We also used Turnitin™ scores from assessments completed by 35 students in the year before we introduced the preprint peer review as a task (2020–2021).

Materials

Preprints. We selected preprints from the field of developmental psychology ((*Bass and Bonawitz, 2022)(*Byers-Heinlein, Gonzalez-Barrero, Schott and Killam, 2024)(*Dotan and Zviran-Ginat, 2022)(*Germain, Gonzalez-Barrero and Byers-Heinlein, 2022)(*Kim, Ahmed and Morrison, 2021)(*Merkley, Sernoskie, Cook, Howard, Draper and Scerif, 2022)(*Peake and Rodríguez, 2020)(*Potter and Casey, 2023)(*Sullivan, Bale and Barner, 2018)(*Vigliocco, Motamedi, Murgiano, Wonnacott, Marshall, Milán-Maillo and Perniss, 2019); (marked with asterisks in the References) that would be comparable and relevant to the students' interests in the module subject. For Cohort 1, we selected six preprints by searching preprints on PsyArXiv using the terms 'language', 'literacy', 'numeracy', 'play', 'media' and 'schooling', filtering these to studies within the early childhood and infancy category, and selecting six through discussion among the authors. For Cohort 2, we selected four preprints (due to a smaller cohort size) through discussion between JM and VK regarding their suitability for the module.

Review rubric. We developed a review rubric with 41 questions that targeted different aspects of a preprint (summary, study design, sampling, analyses, transparency, rigour and impact), which was operationalised as a Qualtrics survey. Questions were developed in iterative discussions amongst a team of academics including all authors and several others who were not involved in this study. The rubric reflects consensus on information that is important to consider when vetting the scientific rigour of preprints. Twenty-six rubric questions required extraction of relevant information from the preprint and 15 were evaluative in nature (see Supplemental Appendix 1 for the list of questions). For each rubric question, we asked how easy it was to answer that question using a visual analogue slider scale anchored at 0 'extremely difficult' and 100 'extremely easy', with responses measured in invisible +1 increments. The full rubric can be accessed here: https://bit.ly/scibehrubric_qualtrics or from the Open Science Framework (Holford, Kempe, McLean, Holcombe and Puebla, 2023).

Procedure

Each student selected one preprint out of the assigned set for their cohort on a first-come-first-serve basis, such that each preprint had approximately 10 students assigned to it. Students were tasked to write a review of their assigned preprint of up to 1,000 words. They were told that the review should include a one-paragraph summary of the reviewed study and elaborate on their critiques of the research and suggestions for improvement. To aid them in this task, students attended a lecture that explained the peer-review process and a tutorial that provided support for the assessment by introducing the rubric and its questions. Students were told that using the rubric would aid them in the process of peer review but were not explicitly instructed to incorporate rubric items into their writing. The rationale was to scaffold the reviewing process without explicitly structuring their writing and to allow students to select rubric items if relevant as a basis for evaluative judgement. In other words, the rubric aimed to give students ideas to think about when reviewing research. Students' rubric responses were not graded, but they were required to submit them to receive 20% of their assignment mark. Students were given the rubric via Qualtrics software, which always presented the rubric questions in the same order, but students could return to earlier questions in the rubric using a back button. They could also pause and return to complete an unfinished rubric at a later stage. All questions in the rubric were optional, meaning that students could submit an incomplete rubric to receive their rubric mark.

Students were also provided with anonymised samples of historical peer reviews that experienced psychology researchers had written for academic journals prior to this study (i.e. they were completed without use of the rubric for different manuscripts), obtained from these reviewers' own records. To improve training further, students in Cohort 2 also had an extra session where they completed a Delphi-method discussion designed to help them with judgements of replicability (described in Pearson et al., 2021). In the Delphi session, the Cohort 2 students first completed eight questions regarding the credibility of the preprint, following which they were divided into groups to discuss their answers, and then they answered the same questions again. This addition to the training was provided to aid data gathering of the RepliCATS project (Fraser et al., 2023).

Measures

As part of their completion of the module, participants provided the following data:

1. Completed rubric for the preprint they were reviewing.
2. Ratings of how easy it was to answer each rubric question.
3. Reviews written for their chosen preprint, which were submitted through the Turnitin™ system that checks and scores the submission for similarity to other written sources on the internet and on the Turnitin database of student submissions, after the references have been removed. This process generated the Turnitin™ score.

Analytical approach

For our analysis, we considered (i) the completed rubrics by students (which included responses to the questions and evaluations of how easy it was to answer those questions) and (ii) the reviews that the students subsequently produced as their module assignment. To compare students with academic reviewers, three of the authors (DH, JM, VK) also completed the same task for all of the preprints, following the same process of completing the rubric and then writing a review for each preprint. For the second cohort, a fourth author (AH) also contributed a rubric and review for two of those preprints using the same procedure. Two research assistants coded all written reviews (by students and academics) for instances of each of the 26 rubric items requiring extraction of relevant preprint information. For moderation purposes, research assistants and one of the authors (JM) first coded the same three reviews in each cohort and met to discuss discrepancies before the research assistants were each assigned reviews to code independently. In addition, a subset of the reviews were coded by both research assistants to enable reliability checks. For Cohort 1, 22 (28%) reviews were coded by both research assistants and for Cohort 2, 8 (15%) were coded by both research assistants. On average across all variables and all preprints, there was a good proportion of coding agreement (Cohort 1: 86%, Cohort 2: 77%). Kappas showed substantial agreement among coders for Cohort 1 (kappa = .65) and moderate agreement for Cohort 2 (kappa = .46), but this metric can be overly sensitive for small samples.

For an exploratory (i.e. not pre-registered) analysis, we obtained Turnitin™ scores for the student reviews (i.e. a percentage indicating the similarity of the written content to other sources), for comparison with Turnitin™ scores from the 2020 to 2021 assignment that preceded the preprint-review assessment. Due to technical issues, Turnitin™ scores were not available for three students in Cohort 1, so they are not included in this analysis. This previous assignment was a 1,500-word essay reviewing evidence on a topic of their choice relevant to early years education (e.g. Does outdoor learning improve cognitive development?; Can

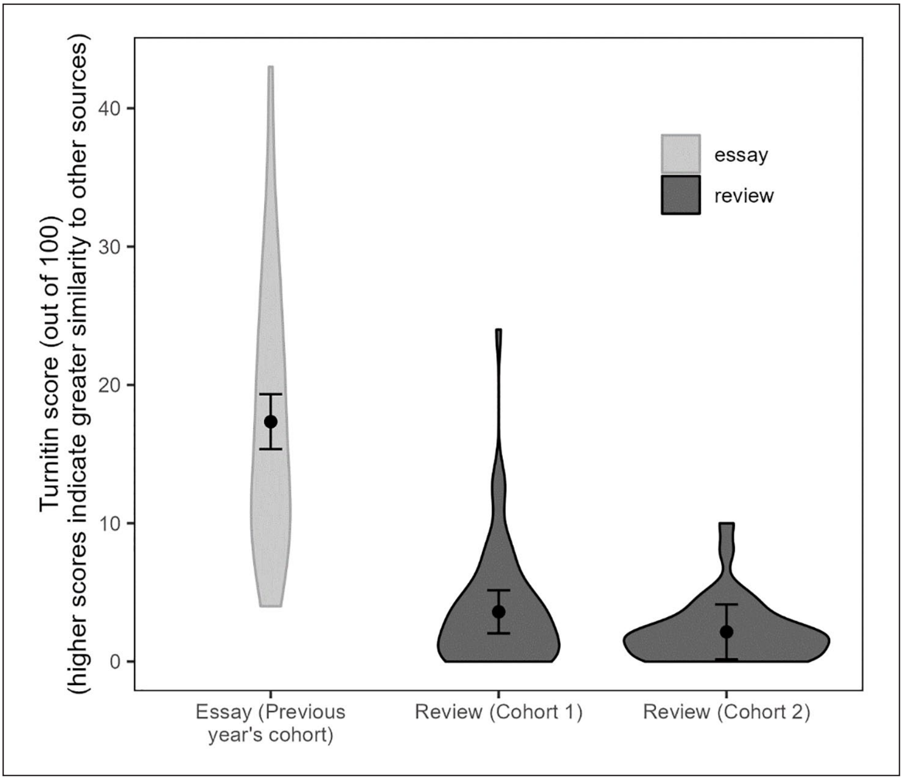

Figure 1. Distribution and estimated marginal means of Turnitin™ scores with 95% confidence intervals for assignments.

educational apps facilitate early reading skills?). Both assignments required students to display similar skills: evaluate evidence and write a critique, albeit in a different format. For all three cohorts, we were also able to obtain students' scores on a different assignment completed in the same year (PowerPoint slides prepared for a 10-minute presentation; identical for all cohorts in all years), to serve as an indication of their individual propensity to paraphrase external sources. All assignments were completed before the widespread use of ChatGPT and associated AI systems, so it is very unlikely that students used such tools or that they would have influenced their Turnitin™ scores.

Results

RQ1: Can preprint peer review serve as an authentic assessment that fosters active involvement with the task in students?

We conducted an exploratory analysis of variance (ANOVA) comparing the Cohort 1 and 2 Turnitin™ scores for the peer-review assessment with the essay assignment completed by the cohort immediately preceding Cohort 1. As shown in Figure 1, students' Turnitin™ scores for the preprint reviews were significantly lower than for the essays written the year before. This main effect across the 3 years was significant, $F(2, 124) = 74.41$, $p < .001$, $\eta^2_P = 0.55$. Post-hoc

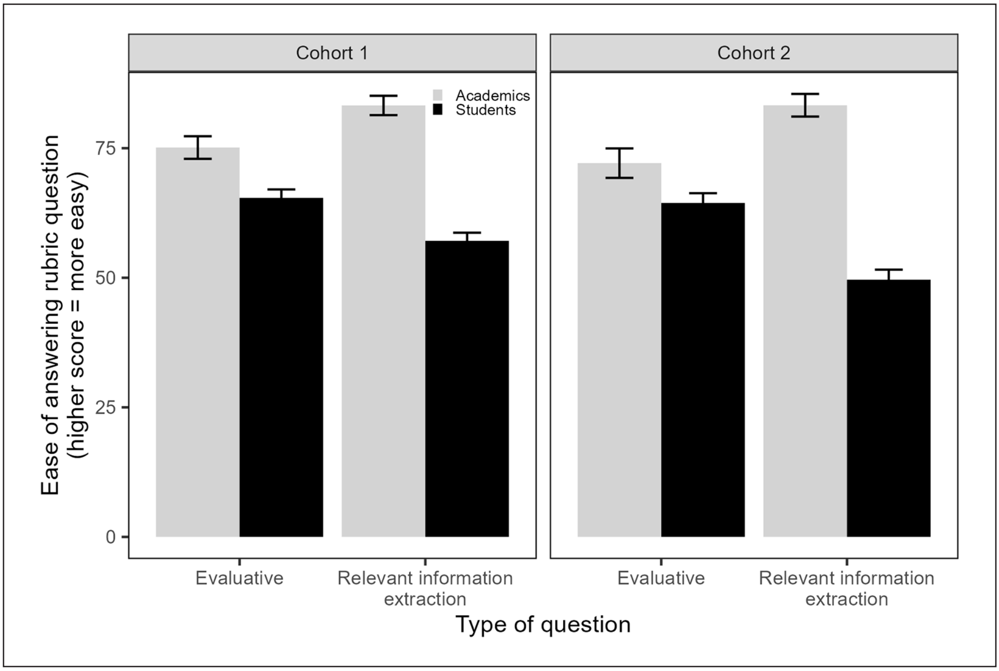

Figure 2. Mean subjective ratings with 95% confidence intervals of how easy it was to answer rubric questions.

Games-Howell tests for unequal variances between years showed significant differences in Turnitin™ scores between the 2020 and 2021 cohort's essays and Cohort 1 preprint review ($t(42.6)=7.99$, $p<.001$); and Cohort 2 preprint review ($t(37.8)=9.11$, $p<.011$). However the test found no significant difference between the two preprint review cohorts ($p=.09$). In order to control for individual students' propensity to paraphrase external sources, we conducted a robustness check that included students' Turnitin™ scores on their second assessment (PowerPoint slides) as a covariate in the model. The main effect remained significant, $F(2, 117)=61.30$, $p<.001$, $\eta^2_P=0.50$.

RQ2: Does the peer review rubric enable undergraduate students to extract relevant information from preprints?

We first compared subjective ratings of how easy it was to answer the rubric questions. From a score out of 100, on average, students rated ease of answering as 57.03 ($SD=16.65$) in Cohort 1 and 48.89 ($SD=18.89$) in Cohort 2. Unsurprisingly, academics found answering rubric questions easier, Cohort 1: 80.68 ($SD=4.55$); Cohort 2: 79.12 ($SD=6.89$). We predicted, and found in a pre-registered paired-samples *t*-test, that students found questions on extracting relevant preprint information (Cohort 1: $M=54.32$, $SD=17.68$; Cohort 2: $M=45.26$, $SD=19.51$) significantly more difficult to answer than evaluative questions (Cohort 1: $M=63.08$, $SD=16.95$; Cohort 2: $M=61.79$, $SD=17.26$), $t(93)=6.76$, $p<.001$, $d=0.72$; see Supplemental Appendix 2 for similar results for completion rates). This was different from the pattern among academics, who found the information extraction questions easier. Due to the low number of academics, we only include their data

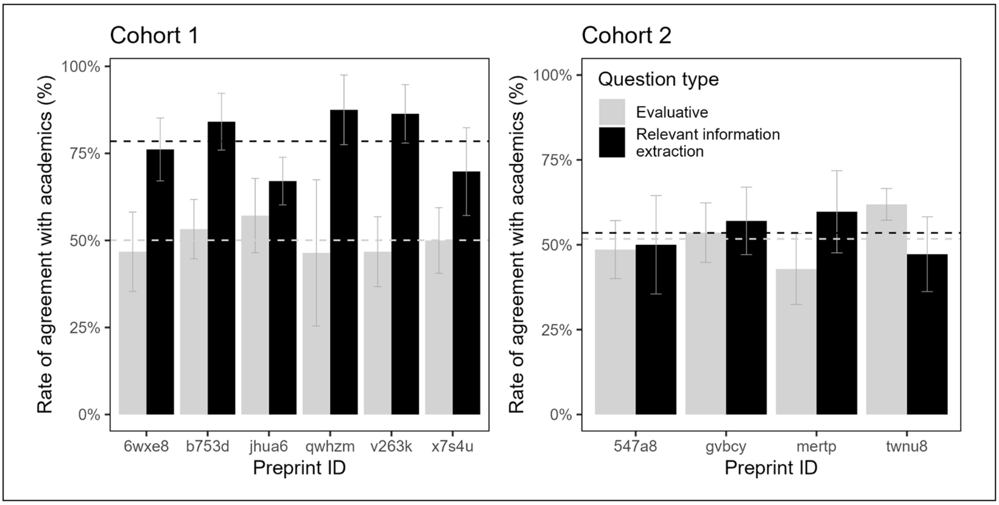

Figure 3. Mean agreement rates, with 95% confidence intervals, between students and academics for information extraction and evaluative questions. Horizontal dotted lines indicated the average agreement for each question type.

for a descriptive comparison with the students and urge caution in interpreting this finding (see Figure 2).

We then assessed the level of agreement between the students' and academics' rubric responses. Although the academics were reasonably well-aligned with each other (overall 79%; 86% for information extraction questions and 67% for evaluative questions), it was sensible to assess students' agreement with academics only for rubric questions where academics were themselves aligned in their responses. To avoid inflating levels of agreement among students and academics, we defined agreement conservatively: we included for this analysis only rubric questions where academics had a modal answer (i.e. a majority agreement) on that question every time it was responded to per preprint associated with that cohort. In other words, we excluded rubric questions as long as there was one preprint for which academics did not have a modal answer (even if they had perfect agreement on this question for the remaining preprints associated with this cohort).

The questions included in this analysis are identified in columns 3 and 4 of the table in Supplemental Appendix 1. Students' rate of agreement with academics was 0.65 (SD=0.12) for the 16 included questions in Cohort 1, and 0.53 (SD=0.15) for the 24 included questions in Cohort 2. As shown in Figure 3, agreement rates were higher for questions requiring relevant information extraction than for evaluation questions, though only marginally in Cohort 2 (Cohort 1: M=0.77, SD=0.17 vs. M=0.5, SD=0.17), $t(59)=8.44$, $p<.001$, d=1.58; Cohort 2: M=0.53, SD=0.19 vs. M=0.52, SD=0.14), $t(35)=0.55$, $p=.585$, d=0.10). We had also pre-registered a linear regression analysis investigating whether subjective ratings of ease of completion were significantly associated with agreement rates, controlling for completion rates as a covariate. This association was not significant (B=0.16, p=.110).

We also assessed whether students were performing significantly above chance in their agreement levels with the academics. We calculated this on an item-by-item basis for multiple-choice items only (reported in Supplemental Appendix 3), since the number of response options determines the item's chance probability. For example, chance agreement would be 50% with two response options, but it would be 25% with four response options. Nine of the 22 multiple-choice

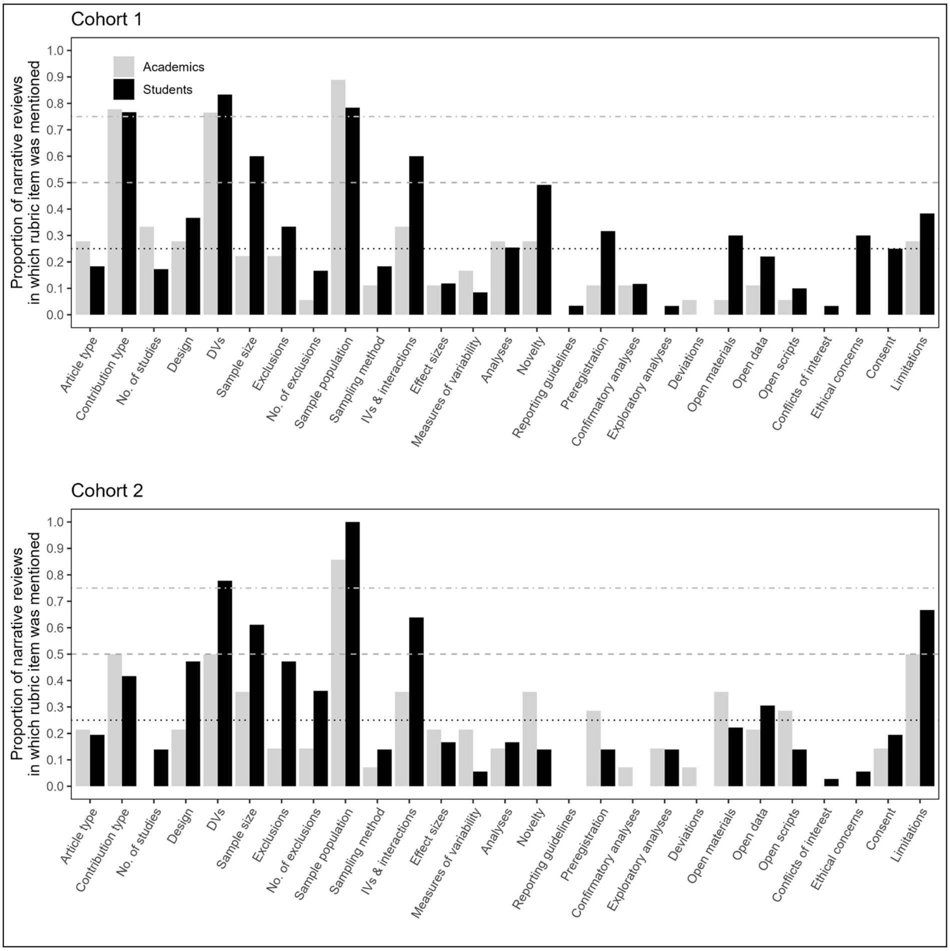

Figure 4. Proportion of students' and academics' written reviews that mentioned each of the rubric items targeting extraction of relevant information. Reference lines are provided at 25%, 50% and 75%.

questions that lent themselves to evaluating agreement showed above-chance agreement in Cohort 1, and 11/22 in Cohort 2. Inspecting the questions that showed above-chance agreement in both cohorts reveals that there was agreement with academics mainly for questions that require a simple check of a fulfilment of a formal criterion, like whether certain reporting guidelines had been adhered to or whether an ethics statement was included. One notable exception is above-chance agreement in the evaluative question E11 (*Rate how likely you think the main result will be to replicate: not very likely/somewhat likely/very likely*), despite variability in the responses.

RQ3: Does the peer review rubric guide students to write preprint reviews of similar content as those by academics?

We first counted how often each of the rubric items that required extraction of relevant preprint information was mentioned in the reviews. For written reviews that were double-coded (i.e. both research assistants coded those), the item was considered as mentioned as long as one of the coders

Table 1. Latent similarity scores and correlations with rubric measures.

	Mean latent similarity (*SD*)	Correlation with number of rubric items mentioned	Correlation with student/academic rubric agreement rate	Degrees of freedom
Cohort 1	0.41 (0.06)	*r* = −.06, *p* = .653	*r* = .04, *p* = .776	58
Cohort 2	0.41 (0.07)	*r* = .09, *p* = .612	*r* = −.06, *p* = .733	34
Overall	0.41 (0.06)	*r* < −.01, *p* = .997	*r* = <.01, *p* = .983	94

considered it to be present in the student's review. Figure 4 shows the proportion of reviews that mention each of the information extraction items. Overall, both the coverage and frequency of rubric items mentioned was greater in students' reviews than in academics' reviews. Students mentioned more of the rubric items ($M=7.86$, $SD=3.52$, compared to $M=6.06$, $SD=2.94$ for the academics), and each rubric item was mentioned in a greater proportion of student reviews ($M=29\%$, $SD=25\%$) than academic reviews ($M=23\%$, $SD=22\%$).

To quantify whether students' written reviews were comparable to academics', we conducted a Latent Semantic Analysis using the lsa-package, version 0.73.3, in R (Wild, 2005). Latent Semantic Analysis computes cosines between vectors of word co-occurrence matrices of a text after dimensionality reduction. We obtained cosines between vectors for each student's review and vectors for each of the academic reviews. For each student, we averaged the cosines of all three comparisons to generate a composite similarity score. If the rubric provides a scaffold for students' review writing to resemble that of academics we would expect a correlation between LSA-similarity scores and indicators of rubric usage under two assumptions: first, that the rubric was based on items academics themselves regarded as important to review in a preprint and would therefore mention in their reviews, and second, that students who completed the rubric similarly to academics would consider the same issues as academics in their reviews.

Pre-registered regression models examined the associations of student-academic review similarity with two rubric measures. First, we tested whether review similarity was correlated with the number of rubric items mentioned. This correlation was not significant, $r=-.01, p=.997$ (see Table 1 for breakdowns by cohort). Second, we tested whether higher agreement with academics on the rubric was associated with student-academic review similarity. This correlation was also not significant, $r<.01, p=.983$ (see Table 1). An exploratory regression analysis also found no significant interaction between the number of rubric items mentioned and rubric agreement rate as a predictor of review similarity, $b=-0.02$, $SE=0.02$, $p=.153$.

Discussion

This study explored whether preprint review, scaffolded by a rubric, could serve as an authentic assessment that allows peer reviewing to be taught to undergraduate students within the framework of university teaching. As part of this project, we explored to what extent students were able to use our rubric to extract relevant information from preprints and to write reviews that resembled those of experienced academics.

We first observed that, after controlling for individual propensity to paraphrase other sources, both cohorts had significantly lower Turnitin™ scores for the preprint reviews compared to a standard essay task completed the year before. This indicates that students most likely generated text themselves compared to other assessments (at the time of the study, producing text through generative AI was not a viable option). Students' active involvement in writing the reviews may have been supported by the low affordance of the peer review task for passively copying from online sources. Although increasing adoption of open reviews provides available models for reviewing in general (Polka et al., 2018; Ross-Hellauer & Görögh, 2019; 'Transparent peer review for all', 2022), it is unlikely that reviews suitable for the particular preprints in the assignment exist online. Further, although copying from other students would have been a possibility for students to avoid being actively involved in the task, lower Turnitin™ scores also indicate less copying from other students. As such, the low Turnitin™ scores point to the suitability of writing peer reviews for actively involving students in an authentic assessment (i.e., one that serves a function beyond student learning) – which can support active learning, develop critical thinking and writing skills, and help students gain self-efficacy (Greene, 2013).

Not surprisingly, students found answering the rubric items subjectively more difficult than academics, particularly the identification of relevant preprint information. However, despite the subjective perception of difficulty, students referred to the relevant information included in the rubric more than academics did when writing their reviews. We also observed that students rated evaluative items as easier to answer whereas the academics showed the opposite trend, although our small sample of academics precluded statistical analysis. This finding may reflect students' lower awareness of the complexity of knowledge required to evaluate scientific work.

We computed agreement in rubric responses between students and academics for those items where academics consistently showed agreement among themselves. For both cohorts, there was agreement in over 50% of these items. There was no evidence that agreement was higher for rubric items that students found easier to answer. Rather, agreement was highest for information extraction items that involved checking whether some reporting criterion (e.g. mention of ethical approval or of participant consent) was fulfilled, especially in Cohort 1. Students also showed above-chance agreement in their evaluation of study replicability. This outcome suggests that undergraduate students are likely to perform well in extracting information about the fulfilment of certain reporting criteria, but may require more guidance to extract complex methodological information. In the future, this rubric may thus serve as a tool for evaluating the efficacy of student training in the identification of a range of methodological features of scientific studies.

We also found that students were more likely to mention rubric items in their reviews than academics, both in breadth (i.e. coverage of items across the rubric) and depth (i.e. frequency of discussing rubric items). This suggests that students used the rubric as a writing guide more than academics did. For example, students were more likely to mention the availability of ethical approval whereas academics tended to mention this only if there were problems. In fact, neither the number of rubric items mentioned in the review nor the degree of agreement with academics' rubric responses were linked to the semantic similarity between students' and academics' reviews. Thus, the lack of similarity did not arise from students getting the rubric questions wrong but from academic reviews being less likely to include information targeted by the rubric.

This observation of low similarity between students' and academics' reviews and lower use of rubric information in academics' reviews may be interpreted in different ways. It is possible that the rubric we designed does not fully capture information that academics tacitly conventionalise over the course of their careers. Reviewing skills that emphasise the theoretical and methodological complexities of research may be difficult to scaffold for undergraduate students. However, we had pooled methodological and reviewing expertise from researchers beyond the author group, and designed the rubric in an iterative process of consensus-seeking about what is important when vetting preprints. It is quite paradoxical that our own reviews did not reflect the rubric items more. One possibility is that we as academics use an evaluation and writing approach based on the review of journal articles, and are implicitly influenced by journal guidelines for reviewers when writing any type of manuscript review. However, when it comes to vetting and evaluating preprints, it can be important to scrutinise a wide range of criteria pertaining to transparency and reproducibility. The lack of similarity between students' and academics' reviews may indicate that we as academics need to overcome ingrained habits and adapt our reviewing practice when it comes to vetting preprints that are in the public domain.

Practical Implications

Even though the rubric did not facilitate similarity between student and academic reviews, our findings suggest that preprint reviewing has merits as an authentic assessment, and the potential to draw undergraduates into evaluating preprints. As a piece of anecdotal evidence, this feature of the assessment was positively highlighted by an external academic examiner teaching a similar programme, who was contracted by the University to provide quality assurance of student assessments – the examiner was unaware of the inclusion of this assessment in a pre-registered study. The low Turnitin™ scores indicated students' active involvement in the task beyond just copying or paraphrasing others' work. It is possible that the prospect of fulfilling a societal need could dissuade students from copying, or indeed, obtaining AI-generated text in the future. At the time of writing, ChatGPT did not provide coherent reviews of prespecified preprints, but only future practice will show whether this will remain the case.

Furthermore, teaching preprint peer review may benefit not just students but also the wider scientific community and society. While peer review remains a key aspect of the academic publishing process, it is often slow (Huisman & Smits, 2017) and increasingly hard to obtain (Aczel et al., 2021). Already a decade ago, the delay between submission and publication at a journal could take up to 18 months, depending on the discipline (Björk & Solomon, 2013). These delays reflect the scarcity of peer-review capacity, due in part to the increased workload of academics (Albert et al., 2016; Fox et al., 2017). The situation is exacerbated by a lack of institutional incentives (Kaltenbrunner et al., 2022) and the disproportionate concentration of peer review among reviewers from developed countries (Publons, 2018; Vesper, 2018), placing considerable demand on this specific reviewer pool. All these factors lead to recommendations that the reviewer pool should be extended beyond the traditional group of published researchers (e.g. Aczel et al., 2021). Given these systemic constraints, incorporating peer-review training in the undergraduate curriculum may be of wider benefit.

Limitations

A major limitation of this study is the small sample of academic experts who provided comparison reviews. This precluded statistical analyses of these reviews, therefore limiting the generalisability of the findings. This reflects one common challenge in peer review, as recruiting experts to provide reviews for comparison can be time-consuming and costly. The issue of statistical power in relation to the academic review sample used for comparison does not, however, affect the findings with respect to the suitability of rubric-aided preprint peer review as an authentic assessment.

Furthermore, we did not compare the reviews written with the rubric (i.e. Turnitin™ scores; agreement with academic reviews) to reviews written without it. We thus had no baseline measures regarding the rubric's benefit for reviewer training. This is because the study's aim was to assess overall suitability of peer review as an authentic assessment. Future research may explore the suitability of the rubric as a pedagogical tool that scaffolds the process of student preprint peer review using a controlled randomised design.

Another limitation is related to the disciplinary specificity of the rubric, as this had been designed primarily for reviewing preprints in psychology. However, we view our findings as proof of concept for the feasibility of training undergraduate students in aspects of the preprint review process. Together with other promising examples in biology (Otto et al., 2023), we hope that additional disciplines may take this as inspiration for developing similar pedagogical tools according to their specific needs and requirements.

Finally, the measurement of the use of rubric items in the reviews was based on coding carried out by student research volunteers who, albeit being trained, were not experts in the peer-reviewing process. As a consequence, the coding reliability varied across cohorts rendering this particular measure less reliable than would be desired. However, it is likely that any errors were of omission, which would affect students' and academics' reviews in equal measure. The findings of greater use of rubric information in student reviews and the lack of a link between use of the rubric and similarity to academics' reviews are thus likely to hold. In future, automated systems could aid comparisons of reviews against rubric items, providing an avenue to research rubric use at a larger scale.

In sum, the present study introduces the preprint review rubric as a potentially valuable pedagogical tool, and shows its suitability for supporting preprint review as an authentic assessment. It also demonstrates usefulness for involving undergraduate students in the extraction of relevant information from scientific papers, providing an avenue to broaden the base of volunteers who can support the much-needed process of preprint vetting. Crucially, students exercise their critical thinking skills through an active process of writing reviews, endowing them with a valuable transferable skill.

Acknowledgements

We would like to thank Amélie Gourdon-Kanhukamwe, Jay Patel, William Page and Cooper Smout for their contributions to the peer review rubric used in the study, and the SciBeh initiative for organising the workshop that led to the conception of this research. We are also grateful to Martin Corley, who was the External Examiner of the assessments investigated in this study, for valuable feedback and suggestions.

Author note

Author contributions

D.H.: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.
J. M.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing - original draft, and Writing - review & editing.
I. P.: Conceptualization, Methodology, and Writing - review & editing. A.O.H.: Conceptualization, Methodology, and Writing - review & editing.
V. K.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing - original draft, and Writing - review & editing.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: At the time at which the study was conceptualised and conducted, IP was an employee of ASAPbio, an organisation that promotes preprints and preprint review activities in the life sciences.
All other authors declare no competing interests.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: DH is supported by funding from Horizon 2020 grant 964728 (JITSUVAX) but this grant did not provide support for the research, authorship, or publication of this article.

ORCID iDs

Dawn Holford https://orcid.org/0000-0002-6392-3991

Alex O Holcombe https://orcid.org/0000-0003-2869-0085

Iratxe Puebla https://orcid.org/0000-0003-1258-0746

Open science statement

Data, materials, and code for the study and analyses reported in article are publicly available at https://osf.io/wgktu/

Supplemental material

Supplemental material for this article is available online.

Notes

1. This paper changed name and first author in 2023. When the students reviewed it the paper's first author was Gonzalez-Barrero and the title was "Bilingual adjusted vocabulary: A developmentally-informed bilingual vocabulary measure."
2. This preprint has subsequently been withdrawn by the authors citing important weaknesses identified by reviewers during journal submission.

References

References marked with asterisks were preprints selected for students to review during the study.

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*(1), 1–8.

Albert, A. Y., Gow, J. L., Cobra, A., & Vines, T. H. (2016). Is it becoming harder to secure reviewers for peer review? A test with data from five ecology journals. *Research Integrity and Peer Review*, *1*(1), 1–8.

Anderson, K. R. (2020). bioRxiv: Trends and analysis of five years of preprints. *Learned Publishing*, *33*(2), 104–109.

*Bass, I., & Bonawitz, E. (2022). *Early environments and exploration in the preschool years*. OSF. https://doi.org/10.31234/osf.io/twnu8

Björk, B. C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923.

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

*Byers-Heinlein, K., Gonzalez-Barrero, A. M., Schott, E., & Killam, H. (2024). Sometimes larger, sometimes smaller: Measuring vocabulary in monolingual and bilingual infants and toddlers. *First Language*, *44*(1), 74–95. https://doi.org/10.31234/osf.io/x7s4u

*Dotan, D., & Zviran-Ginat, S. (2022). Elementary math in elementary school: The effect of interference on learning the multiplication table. *Cognitive Research: Principles and Implications*, *7*, 101. https://doi.org/10.31234/osf.io/547a8

du Rocher, A. R. (2020). Active learning strategies and academic self-efficacy relate to both attentional control and attitudes towards plagiarism. *Active Learning in Higher Education*, *21*(3), 203–216.

Eckmann, P., & Bandrowski, A. (2023). PreprintMatch: A tool for preprint to publication detection shows global inequities in scientific publication. *PLoS One*, *18*(3), e0281659.

Foster, A., Hindle, S., Murphy, K. M., & Saderi, D. (2021). *Open reviewers reviewer guide*. Zenodo. https://doi.org/10.5281/zenodo.5484087

Fox, C. W., Albert, A. Y., & Vines, T. H. (2017). Recruitment of reviewers is becoming harder at some journals: A test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review*, *2*, 1–6.

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P., Pearson, R., Singleton Thorn, F., Ashton, R., Willcox, A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., & Fidler, F. (2023). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLoS One*, *18*(1), e0274429.

Gaudino, M., Robinson, N. B., Di Franco, A., Hameed, I., Naik, A., Demetres, M., Girardi, L. N., Frati, G., Fremes, S. E., & Biondi-Zoccai, G. (2021). Effects of experimental interventions to improve the biomedical peer-review process: A systematic review and meta-analysis. *Journal of the American Heart Association*, *10*(15), e019903.

*Germain, N., Gonzalez-Barrero, A. M., & Byers-Heinlein, K. (2022). Gesture development in infancy: Effects of gender but not bilingualism. *Infancy*, *27*(4), 663–681. https://doi.org/10.31234/osf.io/jhua6

Greene, K. (2013). The theory of active involvement: Processes underlying interventions that engage adolescents in message planning and/or production. *Health Communication*, *28*(7), 644–656.

Hindle, S., & Saderi, D. (2017). PREreview: A new resource for the collaborative review of preprints. *eLife*. https://elifesciences.org/labs/57d6b284/prereview-a-new-resource-for-the-collaborative-review-of-preprints

Holford, D., Kempe, V., Holcombe, A. O., Puebla, I., Patel, J., Smout, C., & McLean, J. (2021). *Testing of pre-print review rubric on student sample*. [Study preregistration.] Open Science Framework. https://doi.org/10.17605/OSF.IO/X9UTN

Holford, D., Kempe, V., McLean, J., Holcombe, A.O., & Puebla, I. (2023). *Engaging undergraduate students in peer review*. [Dataset.] Open Science Framework. https://doi.org/10.17605/OSF.IO/WGKTU

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics*, *113*(1), 633–650.

Kaltenbrunner, W., Pinfield, S., Waltman, L., Woods, H. B., & Brumberg, J. (2022). Innovating peer review, reconfiguring scholarly communication: An analytical overview of ongoing peer review innovation activities. *Journal of Documentation*, *78*(7), 429–449.

*Kim, M. H., Ahmed, S. F., & Morrison, F. J. (2021). The effects of kindergarten and first grade schooling on executive function and academic skill development: Evidence from a school cutoff design. *Frontiers in Psychology*, *11*, 607973. https://doi.org/10.31234/osf.io/b753d

Kirkham J.J., Penfold N.C., Murphy, F., Boutron, I., Ioannidis, J., Polka, J., & Moher, D. (2020). Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ Open*, *10*, e041849.

Lima, N. W., & Nascimento, M. M. (2022). Not only why but also how to trust science: Reshaping science education based on science studies for a better post-pandemic world. *Science & Education*, *31*(5), 1363–1382.

McDowell, G. S., Fankhauser, S., Saderi, D., Balgopal, M., & Lijek, R. S. (2022). Use of preprint peer review to educate and enculturate science undergraduates. *Learned Publishing, 35*(3). https://doi.org/10.1002/leap.1472

McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. *Higher Education*, *85*(1), 85–101.

*Merkley, R., Sernoskie, E., Cook, C., Howard, S., Draper, C., & Scerif, G. (2022). "We don't have things for counting": An exploration of early numeracy skills and home learning experiences of children growing up in poverty in South Africa. *Journal of Numerical Cognition*, *9*(2), 268–284. https://doi.org/10.31234/osf.io/gvbcy

Munasinghe, B. M., Chapman, C., Hewavitharane, C., Hewawasam, G., & Dissanayakege, T. G. (2022). Investing in the academic writing: Training future reviewers and sustaining efficient and quality peer review. *Cureus*, *14*(10), e30341.

Murdock, T. B., & Anderman, E. M. (2006). Motivational perspectives on student cheating: Toward an integrated model of academic dishonesty. *Educational Psychologist*, *41*(3), 129–145.

Otto, J. L., McDowell, G. S., Balgopal, M. M., & Lijek, R. S. (2023). Preprint peer review enhances undergraduate biology students' disciplinary literacy and sense of belonging in STEM. *Journal of Microbiology & Biology Education*, *24*, e00053-23.

*Peake, C., & Rodríguez, C. (2020). Withdrawn: Bidirectional relation of non-symbolic and symbolic numerical systems in first year of kindergarten: the mediating role of ordinality during number learning. *PsyArXiv*. https://doi.org/10.31234/osf.io/qwhzm

Pearson, R., Fraser, H., Bush, M., Mody, F., Widjaja, I., Head, A., Wilkinson, D. P., Sinnott, R., Wintle, B. C., Burgman, M., Fidler, F., & Vesk, P. (2021). *Eliciting group judgements about replicability: A technical implementation of the IDEA Protocol* [Conference session]. In HICSS (pp. 1–10). https://minerva-access.unimelb.edu.au/bitstream/11343/283365/2/0046.pdf

Polka, J. K., Kiley, R., Konforti, B., Stern, B., & Vale, R.D. (2018). Publish peer reviews. *Nature*, *560*, 545–547

*Potter, C. E., & Casey, L. W. (2023). Frequent vs. infrequent words shape toddlers' real-time sentence comprehension. *Journal of Child Language*, 1–11. https://doi.org/10.1017/S0305000923000387

Publons. (2018). *Global state of peer review report*, https://publons.com/community/gspr

Puebla, I. (2022). *ASAPbio toolkit: Setting up your own crowd preprint review activity*. Zenodo. https://doi.org/10.5281/zenodo.5841435

Ross-Hellauer, T., & Görögh, E. (2019). Guidelines for open peer review implementation. *Research Integrity and Peer Review*, *4*, 4.

Sever, R., Roeder, T., Hindle, S., Sussman, L., Black, K. J., Argentine, J., Manos, W., & Inglis, J. R. (2019). bioRxiv: The preprint server for biology. *bioRxiv*, *833400*. https://doi.org/10.1101/833400

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, *70*, 101030.

*Sullivan, J., Bale, A., & Barner, D. (2018). Most preschoolers don't know most. *Language Learning and Development*, *14*(4), 320–338.

Transparent peer review for all. (2022). *Nature Communications*, *13*, 6173. https://doi.org/10.1038/s41467-022-33056-8

Vesper, I. (2018). Peer reviewers unmasked: Largest global survey reveals trends. *Nature*, *2018*, 7–8.

*Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C., Milán-Maillo, I., & Perniss, P. (2019, July 24–27). Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues in their communication to children? In: Goel A., Seifert C., & Freksa C. (Eds.), 41st Annual Conference of the Cognitive Science Society, Montreal, Canada. (pp. 1171–1177). Cognitive Science Society. https://doi.org/10.31234/osf.io/v263k

Warne, V. (2016) Rewarding reviewers: Sense or sensibility? A Wiley study explained. *Learned Publishing*, *29*, 41–50.

Wild, F. (2005): *lsa: Latent semantic analysis*. R package version 0.57.

Author biographies

Dawn Holford is a behavioural science researcher with an interest in the psychology of communication, particularly how people process social and linguistic information when making decisions. Her background is in experimental psychology. She works with the SciBeh initiative to develop infrastructure for better science communication.

Janet McLean is an experimental psychologist who researches the science of learning. Janet is interested in the cognitive processes involved in learning; her research focuses on how they can be integrated into teaching to develop learning skills.

Alex O Holcombe has been involved in initiatives to improve scholarly communication, such as the MetaROR metascience platform, the journal *Advances in Methods and Practices in Psychological Science*, and the Free Journal Network. He co-created the tenzing.club web app to facilitate reporting author information.

Iratxe Puebla is a director of Make Data Count, an initiative that promotes responsible data metrics to enable evaluation of the usage and reach of open data. Iratxe is passionate about open science; in her prior role at ASAPbio, she led projects in support of preprints and transparency in peer review.

Vera Kempe is a professor of Psychology of Language Learning at Abertay University conducting research on first and second language learning, language evolution and variation, as well as emotion and communication. She is interested in teaching and assessment methods that involve students in open access science dissemination and evaluation.