Incorporating Training, Self-monitoring and AI-Assistance to Improve Peer Feedback Quality

Ali Darvishi
a.darvishi@uq.edu.au
The University of Queensland
Brisbane, QLD, Australia

Hassan Khosravi
h.khosravi@uq.edu.au
The University of Queensland
Brisbane, QLD, Australia

Solmaz Abdi
solmaz.abdi@uq.edu.au
The University of Queensland
Brisbane, QLD, Australia

Shazia Sadiq
shazia@itee.uq.edu.au
The University of Queensland
Brisbane, QLD, Australia

Dragan Gašević
dragan.gasevic@monash.edu
Monash University
Melbourne, VIC, Australia

ABSTRACT

Peer review has been recognised as a beneficial approach that promotes higher-order learning and provides students with fast and detailed feedback on their work. Still, there are some common concerns and criticisms associated with the use of peer review that limits its adoption. One of the main points of concern is that feedback provided by students may be ineffective and of low quality. Previous works supply three explanations for why students may fail to provide effective feedback: They lack (1) the ability to provide high-quality feedback, (2) the agency to monitor their work or (3) the incentive to invest the required time and effort as they think the quality of the reviews are not reviewed. To help mitigate these shortcomings, this paper presents a complementary peer review approach that integrates training, self-monitoring and AI quality-control assistance to improve peer feedback quality. In particular, informed by higher education research, we built a set of training materials and a self-monitoring checklist for students to consider while writing their reviews. Also, informed by work from natural language processing, we developed quality control functions that automatically assess feedback submitted and prompt students to improve, if necessary. A between-subjects field experiment with 374 participants was conducted to investigate the approach's efficacy. Findings suggest that offering training, self-monitoring, and quality control functionalities to students assigned to the complementary peer review approach resulted in longer feedback that was perceived as more helpful than those who utilised the regular peer review interface. However, this complementary approach does not seem to affect students judgement (leniency or harshness) or confidence in grading. Directions are suggested to further evaluate and refine peer review systems.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**.

KEYWORDS

Peer feedback, peer review, crowdsourcing in education, learnersourcing, feedback literacy

ACM Reference Format:
Ali Darvishi, Hassan Khosravi, Solmaz Abdi, Shazia Sadiq, and Dragan Gašević. 2022. Incorporating Training, Self-monitoring and AI-Assistance to Improve Peer Feedback Quality. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3491140.3528265

1 INTRODUCTION

Engaging students in peer review has been demonstrated to have various benefits for the reviewers, revieweees and instructors. Reviewers gain the opportunity to develop the ability to better evaluative quality of work (evaluative judgement) [64, 83] and gain experience in providing constructive feedback [54]; reviewees gain the opportunity to receive more frequent and timely feedback [48] from diverse perspectives [16, 70]; and instructors may receive a reduced marking load, which gives them the opportunity to increase student enrollment [36] or reinvest their time more optimally towards enhancing student learning.

Despite these benefits, two fundamental points of critique undermine the reliability of using peer review as an assessment instrument. The first point is related to students' level of knowledge and their lack of ability to differentiate good work from bad as well as an instructor can. Utilisation of rubrics [28, 29], exemplars [13], and comparative judgement where students choose the 'better' of two pieces of works [10, 67] have been shown to be an effective method for helping students develop evaluative judgement [40, 83]. The second point, which this paper aims to contribute to addressing, relates to students failure in providing high-quality feedback, which leads to substantial negative consequences such as lowering standards [94], reducing trust in the outcome [11], and making reviewees less likely to revise their work [81].

There are multiple explanations for why students may fail to provide effective feedback, which our proposed approach aims to address. One explanation is that students are not equipped with the required skill set to provide effective feedback. This is not surprising as many students may never have been asked to provide formal feedback. Indeed, this explanation is supported by a meta analysis of 58 studies that demonstrate providing training for reviewers is one of the most effective ways of improving the quality

L@S '22, June 1–3, 2022, New York City, NY, USA

ACM ISBN 978-1-4503-9158-0/22/06.
https://doi.org/10.1145/3491140.3528265

of peer review [51]. As such, one of the strategies incorporated by our approach is to provide training on how to deliver effective feedback. In particular, we use the guidelines recommended by [12] to develop a set of tips with examples on writing effective feedback. An alternative explanation is that students may be aware of best practices for writing effective feedback, but may not have the agency to monitor their own work to ensure their following of these best practices. Many studies from the field of self-regulated learning have demonstrated that students benefit from strategies that help them monitor their work and regulate their learning [33]. Here, we incorporate a self-monitoring checklist, as one of the well-studied strategies from the self-regulation literature [55, 97], to allow students to track whether their provided comment follows the best practices for writing effective feedback. A third explanation is that students may lack the required incentive to provide feedback diligently [79]. A likely reason is that they think their contributions are not reviewed by instructors. They may therefore put minimal effort in terms of providing feedback [85]. To address this potential challenge, informed by the literature on the use of natural language processing to evaluate the quality of a review [23, 62], we develop quality control functions that automatically assess the quality of the submitted feedback and ask students to improve, if necessary.

We hypothesise that the majority of the students would engage with our approach (H1) and that it enables them to provide higher quality feedback (H2). Also, we hypothesise that a deeper level of engagement with the feedback would make the students less lenient and more confident in their reviews (H3). To test our hypotheses, we conducted a between-subjects field experiment with 374 participants using a learnersourcing platform [43] in which students create learning resources that are evaluated through a peer review process. In our experiment, the control group used the regular interface of the platform for evaluating learning resources in which students first complete a rubric and then provide open-ended feedback to justify their rubric ratings and final decision. The experiment group utilised the same interface with the addition of our complementing approach for providing training material, self-monitoring checklist, and quality-control functions.

This study contributes to the literature by: (1) Integrating multiple approaches to address various causes for students' inability to offer effective feedback, such as a lack of skill, agency, or incentive; (2) presenting cutting-edge AI-assistance techniques to automatically analyse the quality of submitted feedback and notify students of areas where there is room for improvement; and (3) conducting an *in-field* between-subject study on the impact of interventions on feedback quality, which has implications on implementation and adoption of feedback tools.

2 RELATED WORK

Here, we first present a brief review of the literature on systems that support peer review and feedback. We then review some of the existing literature on the use of training, self-monitoring and AI-assistance in higher education, particularly in providing feedback.

Peer review and feedback. The results of prior research comparing the impacts of instructor-led feedback with peer review feedback have suggested that peer review feedback promotes a higher level of learning in students compared to instructor-led feedback [24, 53, 58]. Several studies have also investigated the impact of engaging students in peer review activities on their performance and learning. The results of these studies have shown that engaging students in peer review activities motivates higher levels of student involvement, enhances evaluative judgment, provides a natural environment for communication development, and helps authors better grasp reader demands through the interaction that the peer review process fosters among students [8, 32, 52]. Successful examples of engaging students in peer review activities range from involving more experienced learners to help novices with hints and reviews (e.g., [27]) to pairing students to assess each other's activities (e.g., [80]) or to flag an activity to be further assessed by instructors (e.g., [89]). In order to help students or instructors during the peer review process, different strategies have been implemented in a number of learning platforms. For example, PeerScholar is a web-based platform that provides a viable peer evaluation procedure to help teachers manage writing and critical thinking assessments, as well as student assignment results in a large class setting [69]. Another example is Mechanical TA, an automated peer review system, that aims to advance review quality by involving teaching assistants to evaluate reviews of novices and spot check that of experienced students [91]; Dear Beta and Dear Gamma are two web applications that engage students in peer review activities by enabling them to create hints on their own works and that of their peers [27]; Aropä is an online system that facilitates peer review activity by allowing students to upload assignments, write reviews on peer submissions and view the feedback given on their own works [71]; CrowdGrader enables students to submit, review, and grade homework, as well as receive feedback on the quality of their assignment and reviews [20]; edX, a MOOC platform, pairs students randomly to review their submissions in a peer assessment system to facilitate education in tasks such as writing and design, which are challenging to assess automatically [80]; Peergrade, a web-based peer assessment tool, attempts to improve the feedback quality by an intelligent allocation of reviewers and automatic flagging for instructor moderation [89]. Peer evaluation is also used alongside the content creation in the learnersourcing platform used in our study (RiPPLE) to control the quality of the student-generated resources [28, 29]. Despite the advantages of peer review, the reliability of systems that rely on student judgement for assessment is often critiqued as the quality of work of students has been reported to be quite diverse ranging from very high to very low [3, 7, 22, 26, 82, 87], which limits their adoption. Here, we review recent research that attempted various approaches to help improve peer review, which helped inform the design of our proposed complementary peer review system.

Training. One approach for assisting students in providing high-quality feedback is to provide students with training. For example, Cambre et al. [10] proposed using scaffolding comparison with curated examples to assist students in providing better peer reviews. In addition to using a general rubric, Paré and Joordens [69] incorporated a set of training material including original reading material, the student's own answer, and the abstract and critical thinking guidelines within the PeerScholar platform to assist students during the peer review process. Kulkarni et al. [47] provided training for students during the peer review process by offering them feedback about their bias, standard adaptable feedback texts, and indicating crucial items. Previous controlled studies of peer review training have shown that providing students with training enables them to deliver peer review feedback that is of higher quality than that submitted by untrained students [15].

Self-monitoring. Self-monitoring techniques allow students to be directly involved in the assessment of their own work. As [77] elaborates, self-monitoring/assessment improves students evaluative judgment and enables them to self-regulate their work, which in turn leads to sustainable learning. Self-monitoring techniques are recognised as an effective self-regulation strategy [98]. For example, they have been used successfully in math home work with positive linear trend in self-regulation [76], online learning environments to help note taking and improving achievements [37], and Massive Open Online Courses (MOOCs) to facilitate self-monitoring for self-directed learning [95]. There are also explicit examples of studies that have incorporated self-monitoring within the peer review process to improve the quality of peer feedback. For example, Kulkarni et al. [48] proposed using scaffold comments within PeerStudio to enable students to re-review their peer feedback before its final submission.

AI-assistance. In recent years, there has been an increasing trend in the use of natural language processing techniques (NLP) in different educational setting. In particular, in the field of writing analytics, a sub-domain of learning analytics that concerns utilising analytic techniques for developing a better understanding of writing in educational setting, NLP techniques have been successfully used to automatically analyse the students' writing products and support it through providing personalised automatic feedback [77]. In line with this trend, some recent studies have also offered using AI-assistance approaches and in particular NLP approaches for enhancing the quality of peer review feedback. For example, Xiong et al. [93] proposed a combination of NLP techniques and machine learning to automatically identify a lack of useful features in peer review feedback. Krause et al. [46] proposed using an NLP approach that automatically analyses the feedback language and extracts feedback text features such as specificity and sentiment. Jia et al. [35] proposed another NLP-based approach that leverages the BERT and DistilBERT [74] language representation models to evaluate the quality of peer review comments. The results of various user studies conducted by these researches have demonstrated the efficacy of their approaches in improving the quality of student-generated feedback.

Much effort has also gone into identifying various features of quality feedback such as length and detailedness, scope, alignment to the content, specificity or problem localisation, suggesting solutions, and affective language [14, 34, 45, 63, 93, 99]. On the other hand, most initiatives to help students provide effective peer feedback were confined to a specific feature or method. However, as Henderson et al. [31] implied, achieving effective feedback is challenging, which requires addressing a variety of conditions. Their framework emphasised the significance of synergistic interactions and the various ways to satisfy conditions required for effective feedback. This framework and previous work findings inspired our proposed complementary peer review system, integrating the advantages of three different approaches.

3 METHODOLOGY

The aim of this study is to investigate how the peer review process, outcome and quality are impacted by the addition of a complementary approach that provides training material, a self-monitoring checklist, and quality-control functions compared to the regular peer review. The following research questions guide our investigation into the three hypotheses presented in the introduction:

RQ1: To investigate the first hypothesis about students engagement (H1), we consider this research question: *To what extent do students engage with the complementary approach?*

RQ2: This research question is being considered in order to test the second hypothesis concerning the influence on feedback quality (H2): *What is the impact of the complementary approach on peer feedback quality?*

RQ3: The third hypothesis about influences on peer review decision and confidence rating (H3) is evaluated by the following research question: *What is the impact of the complementary approach on student judgement?*

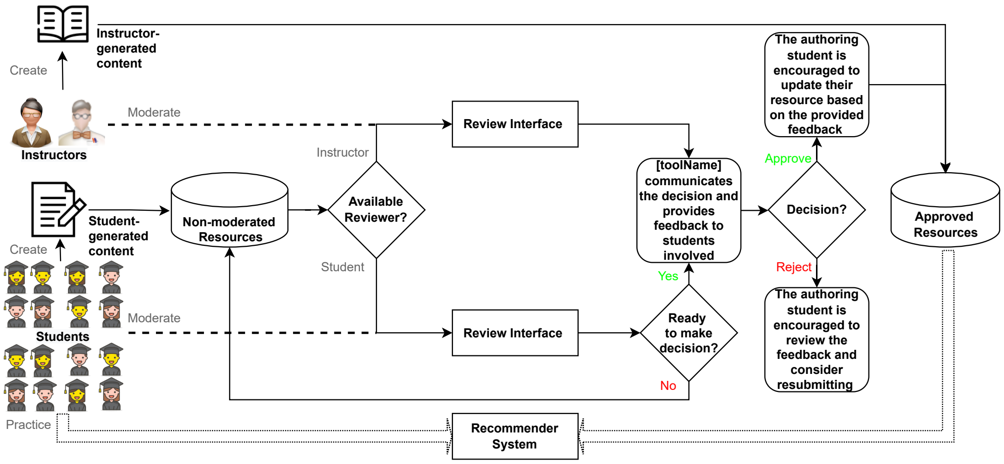

Instructor-generated content
Create
Instructors
Moderate
Review Interface
Instructor
The authoring student is encouraged to update their resource based on the provided feedback
Approve
Student-generated content
Create
Non-moderated Resources
Available Reviewer?
[toolName] communicates the decision and provides feedback to students involved
Decision?
Approved Resources
Student
Yes
Reject
Students
Moderate
Review Interface
Ready to make decision?
The authoring student is encouraged to review the feedback and consider resubmitting
No
Practice
Recommender System

Figure 1: Overview of Processes in RiPPLE

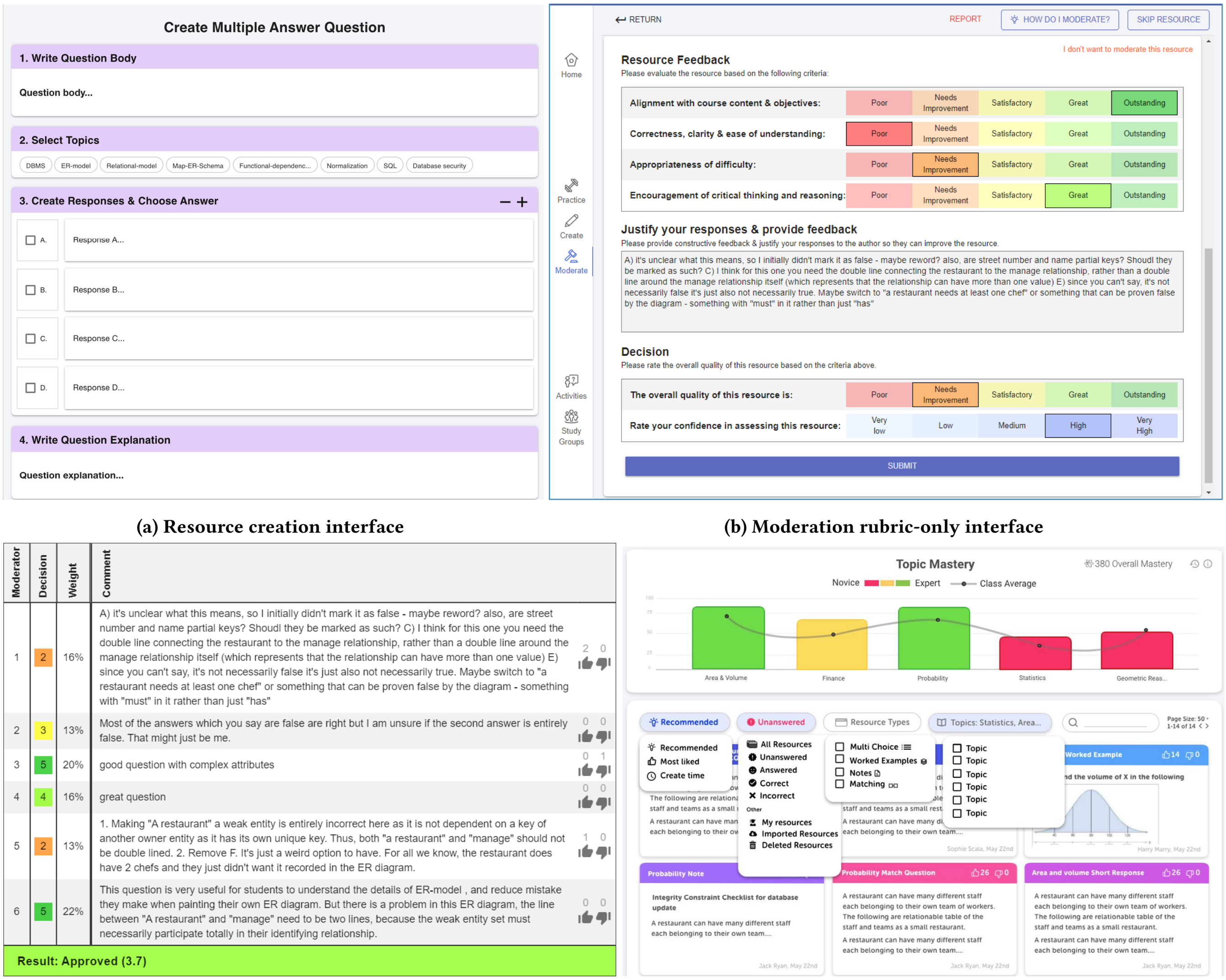
In what follows, Section 3.1 introduces RiPPLE, which is the tool used for conducting the study. Section 3.2 presents the proposed complementary approach for peer review. Finally, Section 3.3 presents the experimental design of the study to address the above-mentioned research questions.

3.1 The Tool: RiPPLE System

At its core, RiPPLE is an adaptive educational system that dynamically adjusts the level or type of instruction based on individual student abilities or preferences to provide a customised learning experience [41]. To provide customised learning for students with different knowledge states, adaptive educational systems require large repositories of learning resources, which are commonly created by domain experts [4]. Such systems are therefore expensive to develop and challenging to scale. Instead of relying on domain experts as developers, RiPPLE uses a learnersourcing approach to engaging students in the creation and evaluation of a range of learning resources [39, 44]. Fig. 1 provides an overview of the main processes in RiPPLE– content creation, peer review, peer review outcome, and practice processes, which are discussed below.

Create. Both students and instructors can create learning resources in RiPPLE [49]. Fig. 2a illustrates the interface used for creating multiple answer questions. Users can create different types of resources, including multiple-choice questions, multi-answer questions, worked examples as well as open-ended notes.

Peer Review. A resource created by an instructor is directly added to a repository of approved resources, whereas a resource created by a student needs to go through a peer review process, also called the



Create Multiple Answer Question

1. Write Question Body

Question body...

2. Select Topics

DBMS · ER-model · Relational-model · Map-ER-Schema · Functional-dependenc... · Normalization · SQL · Database security

3. Create Responses & Choose Answer — +

□ A. Response A...

□ B. Response B...

□ C. Response C...

□ D. Response D...

4. Write Question Explanation

Question explanation...

(a) Resource creation interface

Resource Feedback

Please evaluate the resource based on the following criteria:

Alignment with course content & objectives:	Poor	Needs Improvement	Satisfactory	Great	[✓ Outstanding]
Correctness, clarity & ease of understanding:	[✓ Poor]	Needs Improvement	Satisfactory	Great	Outstanding
Appropriateness of difficulty:	Poor	[✓ Needs Improvement]	Satisfactory	Great	Outstanding
Encouragement of critical thinking and reasoning:	Poor	Needs Improvement	Satisfactory	[✓ Great]	Outstanding

Justify your responses & provide feedback

Please provide constructive feedback & justify your responses to the author so they can improve the resource.

A) it's unclear what this means, so I initially didn't mark it as false - maybe reword? also, are street number and name partial keys? Shoudl they be marked as such? C) I think for this one you need the double line connecting the restaurant to the manage relationship, rather than a double line around the manage relationship itself (which represents that the relationship can have more than one value) E) since you can't say, it's not necessarily false it's just also not necessarily true. Maybe switch to "a restaurant needs at least one chef" or something that can be proven false by the diagram - something with "must" in it rather than just "has"

Decision

Please rate the overall quality of this resource based on the criteria above.

The overall quality of this resource is:	Poor	[✓ Needs Improvement]	Satisfactory	Great	Outstanding
Rate your confidence in assessing this resource:	Very low	Low	Medium	[✓ High]	Very High

SUBMIT

(b) Moderation rubric-only interface

Moderator	Decision	Weight	Comment	
1	2	16%	A) it's unclear what this means, so I initially didn't mark it as false - maybe reword? also, are street number and name partial keys? Shoudl they be marked as such? C) I think for this one you need the double line connecting the restaurant to the manage relationship, rather than a double line around the manage relationship itself (which represents that the relationship can have more than one value) E) since you can't say, it's not necessarily false it's just also not necessarily true. Maybe switch to "a restaurant needs at least one chef" or something that can be proven false by the diagram - something with "must" in it rather than just "has"	2 👍 0 👎
2	3	13%	Most of the answers which you say are false are right but I am unsure if the second answer is entirely false. That might just be me.	0 👍 0 👎
3	5	20%	good question with complex attributes	0 👍 1 👎
4	4	16%	great question	0 👍 0 👎
5	2	13%	1. Making "A restaurant" a weak entity is entirely incorrect here as it is not dependent on a key of another owner entity as it has its own unique key. Thus, both "a restaurant" and "manage" should not be double lined. 2. Remove F. It's just a weird option to have. For all we know, the restaurant does have 2 chefs and they just didn't want it recorded in the ER diagram.	1 👍 0 👎
6	5	22%	This question is very useful for students to understand the details of ER-model , and reduce mistake they make when painting their own ER diagram. But there is a problem in this ER diagram, the line between "A restaurant" and "manage" need to be two lines, because the weak entity set must necessarily participate totally in their identifying relationship.	0 👍 0 👎

Result: Approved (3.7)

(c) Moderation outcome and provided feedback

...nd the volume of X in the following

Integrity Constraint Checklist for database update

A restaurant can have many different staff each belonging to their own team....

A restaurant can have many different staff each belonging to their own team of workers. The following are relationable table of the staff and teams as a small restaurant.

A restaurant can have many different staff each belonging to their own team....

(d) Personalised practice interface

Figure 2: Interfaces of the main processes in RiPPLE

moderation process in RiPPLE [17]. Upon availability of a student to peer review a resource (i.e., the student goes on the moderation tab on the platform), RiPPLE selects and presents a non-moderated resource to the student. Resources are generally selected based on a FIFO (first in first out) queue implementation. However, factors related to concurrency issues (due to the availability of multiple moderators at the same time) or conflict of interest (due to strong ties between the moderator and the author) may impact the selection. Fig. 2b displays the review interface used by moderators for evaluating a resource. It includes a rubric of four items, which asks moderators to rate a resource on alignment, correctness, difficulty level, and critical thinking encouragement. Moderators are then expected to justify their decision and provide feedback to the author before submitting their evaluation. Finally, moderators provide a final decision and their confidence in their rating. Upon availability of an instructor to review a resource (i.e., the instructor goes on the moderation tab on the platform), RiPPLE employs spot-checking algorithms [88] to identify and share with the instructor a resource that can benefit the most from expert judgment. Generally, resources that have a high disagreement among moderators or are close to the boundary decision line are good candidates for being checked by instructors. Reviews from instructors are considered final, meaning their decisions are considered the ground truth without considering evaluations from students.

Peer Review Outcome. RiPPLE considers a number of factors to decide whether or not it is ready to make a decision about the quality of the resource under moderation, including the number of peer reviews received, reliability of the reviewers that have submitted a review, which is computed using a probabilistic method inspired by the well-known expectation-maximisation [60], and the level of agreement between the received reviews. If the system is not ready to make a decision, then the status of the resource remains unchanged, and it awaits further moderations. When the system is ready to make a decision, it uses explainable consensus algorithms, as discussed in [18], to update the status of the resource to approved or rejected. The same algorithm is used to update the reliability rating of the author and student moderators of the resource. The authors of approved resources are encouraged to update their resources based on the feedback provided. Their resource is added to a repository of approved resources that are used in the adaptive engine of RiPPLE. The authors of rejected resources can update and resubmit their resource; however, if resubmitted, the resource will be considered a new submission and will have to go through the moderation process again.

Fig. 2c shows an example of how peer review outcome and feedback are shared with author and reviewers. Instructors can only view the names of the student moderators [removed in the figure for privacy protection], decisions which are between 1 (poor) to 5 (outstanding), current reliability ratings, confidence levels as determined by the moderators which are between 1 (very low) to 5 (very high), the weights of contribution towards making the final decision, ratings on the rubrics of four items (i.e., content alignment, correctness, difficulty level, and critical thinking) which are between 1 (poor) to 5 (outstanding), and comments provided by each moderator. The author and student moderators can see the decisions, confidence levels, contribution weights, rubric item ratings and the provided comments; however, they cannot view the identity or the current reliability ratings of the other moderators. They can also mark each feedback as being helpful or not by clicking on likes and dislikes.

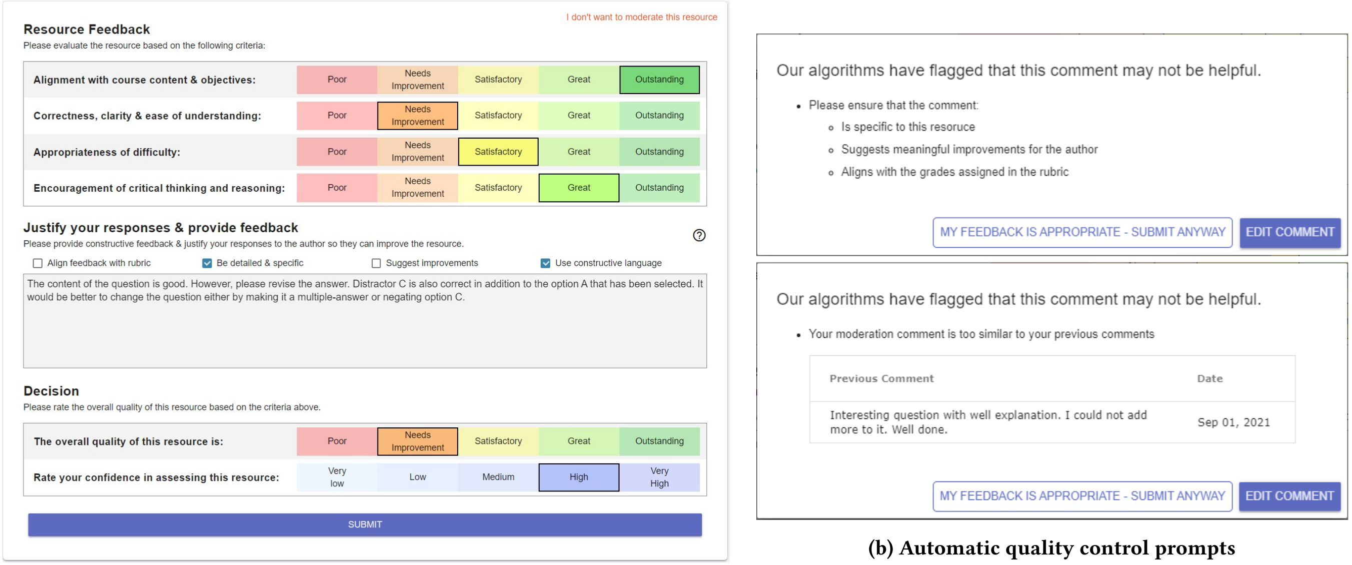
Personalised Practice. Fig. 2d illustrates the interface used for providing personalised practice opportunities for students. The top part of the figure represents an interactive visualisation widget, in form of an open learner model [2, 9], that allows students to view an abstract representation of their knowledge state based on a set of topics associated with a course offering. The colour of the bars, determined by the underlying algorithm modelling the student, categorises competence into three levels. Namely, for a particular unit of knowledge, red, yellow and green signify inadequate competence, adequate competence with room for improvement, and mastery, respectively. Currently, RiPPLE employs an Elo-based rating system for approximating the knowledge state of users with the results translated into coloured bars [1]. The lower part of the screen displays learning content from the repository of approved resources that are recommended to a student based on their learning needs using the recommender system outlined in [38].

3.2 The Complementary Approach for Peer Review

This section describes the design choices and techniques used to complement the current peer review interface, which is illustrated in Fig 2b. Three main strategies of providing training using tips, self-regulation using checklists and automated oversight using NLP functions are employed by the approach. Fig. 3 demonstrates an overview of the various interfaces incorporated by the complementary approach.

Training and Self-regulation. Informed by higher education research on feedback quality [31, 34, 63, 99], a set of training materials, shown in Fig. 9 in Appendix A, are developed to help students in providing constructive and effective feedback. Students are guided to consider four criteria: (1) Be aligned with rubrics, (2) Be detailed and specific, (3) Suggest improvements, and (4) Use constructive language. Explicit positive and negative examples of how the tips can be utilised in practice are included in the training. Also informed by higher education research on self-regulated strategies [68, 98], we incorporated a self-monitoring checklist to reinforce the use of the guidelines provided by the training material and help students track whether they have incorporated the tips in their feedback. As shown in Fig. 3a, the checklist is on top of the text box where students write their comments and the training material is accessible by clicking on the ? button. This approach aims to help students develop feedback literacy and to gain the ability and the agency to regulate their own learning.

Automatic Quality Control. Informed by work from the NLP community [23, 61, 62, 72, 73], we developed a set of quality control functions that automatically assess the quality of the submitted feedback and prompt students to improve the feedback. The first function automatically detects if a suggestion has been expressed in the submitted feedback using an approach adopted from the work by [62]. If no suggestion has been detected, another function measures the relatedness between the provided textual feedback



I don't want to moderate this resource

Resource Feedback

Please evaluate the resource based on the following criteria:

Alignment with course content & objectives:	Poor	Needs Improvement	Satisfactory	Great	[✓ Outstanding]
Correctness, clarity & ease of understanding:	Poor	[✓ Needs Improvement]	Satisfactory	Great	Outstanding
Appropriateness of difficulty:	Poor	Needs Improvement	[✓ Satisfactory]	Great	Outstanding
Encouragement of critical thinking and reasoning:	Poor	Needs Improvement	Satisfactory	[✓ Great]	Outstanding

Justify your responses & provide feedback

Please provide constructive feedback & justify your responses to the author so they can improve the resource.

□ Align feedback with rubric ☑ Be detailed & specific □ Suggest improvements ☑ Use constructive language

The content of the question is good. However, please revise the answer. Distractor C is also correct in addition to the option A that has been selected. It would be better to change the question either by making it a multiple-answer or negating option C.

Decision

Please rate the overall quality of this resource based on the criteria above.

The overall quality of this resource is:	Poor	[✓ Needs Improvement]	Satisfactory	Great	Outstanding
Rate your confidence in assessing this resource:	Very low	Low	Medium	[✓ High]	Very High

Our algorithms have flagged that this comment may not be helpful.

- Please ensure that the comment:
  - Is specific to this resoruce
  - Suggests meaningful improvements for the author
  - Aligns with the grades assigned in the rubric

Our algorithms have flagged that this comment may not be helpful.

- Your moderation comment is too similar to your previous comments

Previous Comment	Date
Interesting question with well explanation. I could not add more to it. Well done.	Sep 01, 2021

(b) Automatic quality control prompts

(a) Self-monitoring checklist

Figure 3: Complementary interface for peer review including: (a) complemented peer review interface, and (b) automatic quality control prompts.

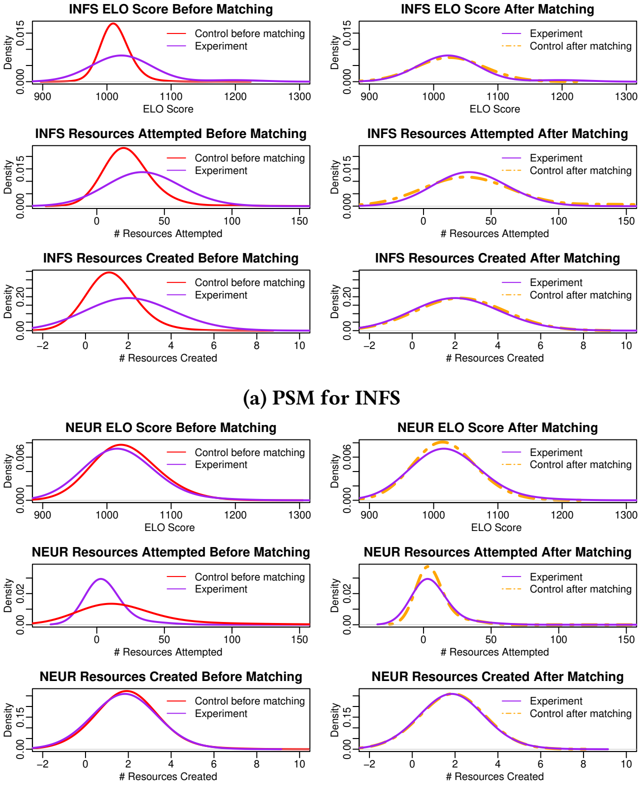
and the resource context. To score the semantic textual similarity of the feedback-resource pair, we used SBERT [73] as the encoder function to calculate the cosine similarity score. The score ranged in $[-1, 1]$ was used as a measure of relatedness between the two representations– feedback and resource. SBERT is developed based on a neural language model called BERT (Bidirectional Encoder Representations from Transformers) [23], which is pre-trained on a large language corpus to encode sentences in the way that similar sentences are close to each other in the embedding space. These models are pre-trained on large amounts of text and proven to have state-of-the-art performance in many NLP tasks with no supervision. Finally, we developed a function that utilises the GLEU (Google's biLingual Evaluation Understudy) measure [92] to calculate the similarity between the current submitted text and previous comments of the moderator using n-grams. This function complements the auto quality control to reduce the chance of gaming the system by submitting the same general comment (e.g., 'Good question, it needs a good understanding of the course content to be solved.') several times for different resources that can pass the first two functions. Fig 3b shows examples of the prompts given by our automatic quality control function to students. Students are prompted to edit their comment based on the provided feedback or to state that they think their feedback is appropriate as is. The aims of this approach were twofold: Part of the intention was to complement the training provided by the other approach and explain why the system believes their provided feedback requires improvement. The other part of the intention was to reduce poor behaviour (lack of effort in providing feedback) by introducing a certain level of risk and oversight that informs students that the quality of their contributions is being monitored.

3.3 Study Design

3.3.1 Data Collection and Experimental Settings. To answer the research questions under investigation, we conducted a between-subject experiment using two consecutive offerings (Semester 1 and Semester 2) of two undergraduate courses in 2021, namely The Brain and Behavioural Sciences (NEUR) and Introduction to Information Systems (INFS) at the University of Queensland.[1] For this study, students enrolled in Semester 1 offering of the courses that used RiPPLE with the non-complemented peer review (NP) interface as shown in Fig 2b were considered the control group (in the remainder of this paper, the control group is referred to as NP). Students enrolled in Semester 2 offering of the courses that used RiPPLE with the complemented peer review (CP) interface (as shown in Fig 3 were considered the experiment group (in the remainder of this paper, the experiment group is referred to as CP). While randomised controlled trials are gold-standard tests for establishing causality in many fields, they are often subject to threats to unethically disadvantage the learning opportunities for students in the educational setting. In addition, the risk of data contamination raises in the face-to-face offerings when using the platform in different experimental settings. This risk increases as control group students from the same course find that other functionalities were offered to their peers in the experiment group. To comply with ethical considerations and decrease the chance of harming the learning opportunities for students in this study, we have taken into account the following considerations. First, only data from the students that had provided their consent in RiPPLE was included in our analysis. However, all users can use the RiPPLE regardless of their response to the consent form. Second, students from the previous semester

[1] Approval from the University Human Research Ethics Committee (#2018000125) was received for conducting the experiment.

were selected for the control group when the complementary approach was not offered in the peer review process of RiPPLE. The two offerings of each of the courses were largely similar. There were no significant differences regarding the program degree in which students were enrolled. Furthermore, the courses were largely unchanged across the two semesters: the same course contents for lectures, tutorials and practical sessions were taught by the same instructors and tutors. Both courses used a rubric in which students' engagement with RiPPLE had a 10% contribution towards students' final grade. The grade associated with RiPPLE in both courses was conditional on students' engagement with the moderation process, but each course had a slightly different requirement; In NEUR students were required to moderate twice as many resources compared to INFS. Also, we apply propensity score matching (PSM) [5] to match each student in the experimental group with a student from the control group so that the two students are similar on a set of their characteristics (covariates). The baseline covariates selected for this study are (a) the knowledge state of users approximated by the Elo-based rating system, (b) the number of resources that a student has engaged with (i.e., attempted), and (c) created.



(a) PSM for INFS

(b) PSM for NEUR

Figure 4: Distribution of propensity score and covariates before and after matching for: (a) INFS, and (b) NEUR.

Distributions of the selected covariates (ELO Score, # Resources Attempted, and # Resources Created) for students in control and experiment groups in INFS and NEUR are shown in Fig 4 before and after matching. They represent the achievement of PSM in reducing the inequalities between the experimental and control groups.

Table 1: Overview of the experimental groups

Course	Group	#Students	#Resources	#Reviews
INFS	NP	70	165	342
	CP	70	216	304
NEUR	NP	117	703	1,290
	CP	117	573	1,247
Total		374	1,657	3,183

Table 1 provides an overview of the control (NP) and experiment (CP) groups in terms of the number of students (#Students), number of resources (#Resources), and number of peer reviews (#Reviews). The peer reviews in both groups were double-blinded; that is, neither the reviewers nor the authors (reviewees) knew the identities of their peers in the review process of each resource. The independent variable of the experiment was the peer review condition, Non-complemented peer review (NP) for the control group and complemented peer review (CP) for the experiment group. Dependent variables included length of comments, helpfulness likes, quality and confidence ratings as further explained in Section 3.3.2.

3.3.2 Metrics and Analysis. Here, we outline the metrics used and analysis performed to answer each of the research questions. We used t-test to perform statistical analysis of the reported results and the Chi-squared test of association for categorical data, where $p < 0.05$ is used as the criterion for assessing statistical significance. We also reported the corrected effect size using Hedges' g_s and common language effect sizes where appropriate, as recommended by Lakens [50], and Cramer's V Coefficient (V) to measure the relative strength of an association between categorical variables.

RQ1: Engagement. For RQ1, we examined the engagement level of students with the complementary approaches by measuring the percentage of: (1) students who accessed the training material, (2) students who made use of the checklists at least once, (3) comments that were flagged for review, and (4) flagged reviews that were revised. Results of this investigation are reported in Section 4.1

RQ2: Impact on Feedback. For RQ2, we examined the impacts of the complementary approach on the provided comments by measuring the: (1) length (word count) of the comments, (2) the percentage of the comments that received at least one helpfulness like from other reviewers of the same resource and (3) the quality of textual feedback by manually coding 10% (i.e., 163 for NP and 154 from CP) randomly selected comments from each group. Figure 2c demonstrates the interface used for capturing helpfulness like of the other peer reviewers. Results of this investigation are reported in Section 4.2

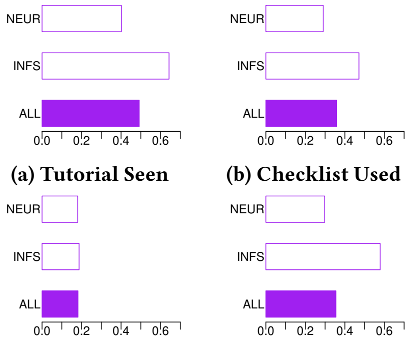
RQ3: Impacts on Judgement For RQ3, we examined the impacts of the complementary approaches on student judgement by measuring the average values of student decision and confidence ratings (shown on Figure 2b). Results of this investigation are also reported in Section 4.3.

4 RESULTS

This section reports the results of our investigation in answering the three RQs proposed in Section 3.

4.1 RQ1: Engagement with the Complementary Approach

Fig 5 shows students' engagement with the complementary approach. Fig 5a shows that 49.2% (i.e., 92 out of 187) of students across the CP group, 64.3% in INFS and 40.2% in NEUR, have accessed the training material. Fig 5b shows that 35.9% (i.e., 67 out of 187 students) of students across the platform, 47.1% in INFS, and 29.1% in NEUR have used the checklists at least once. These results indicate that around half of the students might have explicitly benefited from the training and self-monitoring strategies.



(a) Tutorial Seen (b) Checklist Used

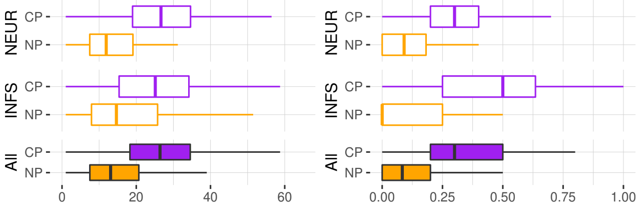
(c) Reviews Flagged (d) Reviews Revised

Figure 5: Percent of students in the CP group who have (a) seen the tutorials and training for providing feedback and (b) used the checklist during the peer review process as well as, probability of (c) comments getting flagged and then (d) being revised before final submission

Fig 5c shows that 18.2% (i.e., 282 out of 1,551) of comments across the platform, 18.8% in INFS, and 18.0% in NEUR were flagged when submitted. However, only 35.5% (i.e., 100 out of 282) were revised among these flagged comments, 57.9% in INFS and 29.8% in NEUR. Fig 5d shows that in the 64.5% of cases when the system flagged a comment, students indicated that they think their feedback is appropriate and does not require any revision.

4.2 RQ2: Impact on Peer Feedback

Comment Length. Fig 6a presents the changes in the length of comments from different experiment groups across all collected data and for each course– INFS and NEUR. As it is indicated, across all data, students in the CP group have provided significantly longer comments (M = 29.2, SD = 16.9) than students in the NP group (M = 15.8, SD = 12.2), $t(372) = 8.77$, $p < .001$, 95% CI [10.34, 16.32], Hedges' g_s = 0.90, 95% CI [0.69, 1.12]). For a randomly selected pair of individuals from two groups, the common language effect size (CL) indicates a 74% chance that the comment length of a person from the experiment group is longer than the comment length of a person from the control group [for calculations for CL, see Lakens [50]].



(a) Comment Length (words) (b) Rate of likes (helpfulness)

Figure 6: Students' effort in providing peer review per each group: (a) average length of comments in words, and (b) percentage of reviews that were perceived as helpful

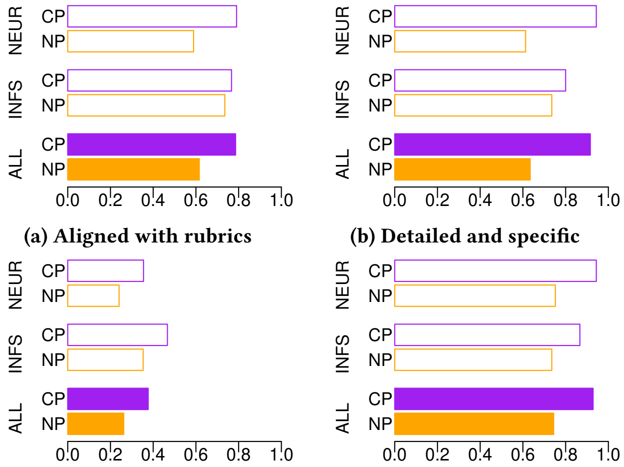
Helpfulness Likes. Fig 6b shows the rate of users' reviews in each group that received at least one helpfulness rating from other reviewers. This analysis reveals that after the review process, the experiment group (M =0.36, SD =0.22) received more likes from peers on the helpfulness of their comments than the control group (M =0.11, SD =0.13), t(372) =13.14 , p<.001, 95% CI [0.21 , 0.28], Hedges' g_s =1.36, 95% CI [1.13, 1.58]. For a randomly selected pair of individuals, the CL effect size indicates an 83% chance that the like per review rate of a person from the experiment group was higher than the like per review rate of a person from the control group.

Coding of Comments. The codebook is based on the four criteria introduced in the training materials (as described in Section 3.2 and Fig ??). One of the authors and an independent researcher coded these randomly selected comments individually to indicate whether the feedback: (1) was aligned with rubrics, (2) was detailed and specific, (3) suggested improvements, and (4) used constructive language, the coders were blind to the conditions. Cohen's kappa coefficient of 0.87 shows an excellent agreement between the coders with the inter-rater agreement of 94.6% across all codes.

Alignment. As shown in Fig 7a, the analysis revealed that 62.6% of peer reviews in the NP group and 78.6% of the CP group had been 'aligned with rubrics' (for INFS: NP= 73.5%, CP=76.7% and NEUR: NP= 58.9%, CP= 79.0%). The Chi-Square Test of independence showed that there was a significant relation between the peer review condition and the alignment, $\chi^2(1, N = 317) = 10.4$, $p = .001$, $V = .18$, the CP group was more likely to align their feedback with rubrics compared to the NP group.

Specific. Fig 7b indicates that the code 'detailed and specific' accounted for 63.4% of the comments in the NP group, compared to 91.6% in the CP group (for INFS: NP= 73.5%, CP=80.0% and NEUR: NP= 61.2%, CP= 94.4%). There was also a significant relationship between these variables, the feedback from the CP group was coded as significantly more detailed and specific, $\chi^2(1, N = 317) = 34.7$, $p < .05$, $V = .33$.

Suggestion. Fig 7c shows that 'suggested improvements' constituted in 26.2% of the coded comments from the NP group compared to 37.6% for the CP group (for INFS: NP= 35.3%, CP=46.7% and NEUR: NP= 24.0%, CP= 35.5%). The association between these variables was significant, $\chi^2(1, N = 317) = 4.6$, $p = .031$, $V = .12$. The CP group was more likely than the NP group to include an explicit improvement suggestion in their feedback.



(a) Aligned with rubrics

(b) Detailed and specific

(c) Suggested improvements

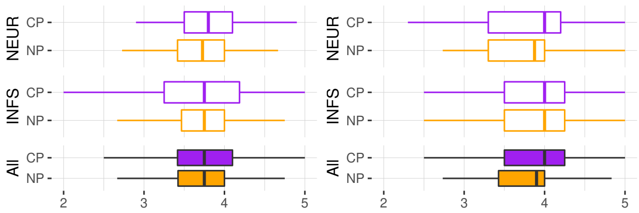
(d) Used constructive language

Figure 7: Percent of students' comments in each group that: (a) were aligned with rubrics, (b) were detailed and specific, (c) suggested improvements, and (d) used constructive language.

Constructive language. Finally, Fig 7d illustrates that there was a significant relation between the peer review condition and the percentage of comments concerning 'constructive language' with 74.3% for the NP group compared to 92.9% for the CP group (for INFS: NP= 73.5%, CP=86.7% and NEUR: NP= 75.2%, CP= 94.4%), $\chi^2(1, N = 317) = 18.7,\ p < .001,\ V = .24$.

4.3 RQ3: Impacts on Judgement

Fig 8a compares the decision ratings from different experiment groups across all of the collected data as well as INFS and NEUR. The reported results suggest that across all of the collected data, there was no significant difference in the decision rating ($t(372) = 1.23$, $p = 0.22$) for the students in the CP group ($M = 3.77$, $SD = 0.56$) and students in the NP group ($M = 3.71$, $SD = 0.49$). The reported results for INFS and NEUR follow a very similar pattern to that of the reported results for the entire data in which there was no significant difference in the decision rating of the students in the NP group and the students in the CP group.



(a) Decision Ratings

(b) Confidence Ratings

Figure 8: Student subjective rating submitted with their peer review on the quality of their peers' work: (a) average decision rating and (b) average self-assessment of confidence,

With regards to confidence in rating, the reported results in Figure 8b suggest that across the entire data there was no significant difference ($t(372) = 1.69$, $p = 0.09$) between students in the CP group ($M = 3.91$, $SD = 0.64$) and the NP group ($M = 3.80$, $SD = 0.57$). For INFS and NEUR, the results were very similar in that there was no significant difference in the confidence in rating for the students in the NP group and students in the CP group.

5 DISCUSSION AND CONCLUSION

This paper contributes to the growing literature on the development of methods and practices for the effective adoption of peer review in higher education. Employing insights and best practices from feedback literacy, self-regulation, and natural language processing research, we proposed a novel complementary peer review approach that incorporates training, self-monitoring, and oversight techniques for students to consider while writing their reviews. We evaluated our approach using a between-subject experiment with 374 participants. Key findings are highlighted below.

RQ1. Engagement with the intervention. The findings presented in Section 4.1 did not support our initial hypothesis (H1) that the majority of students would utilise the complementary approach. According to the results, roughly a half of the students who had access to the training materials and the checklist engaged with them explicitly. Although some additional students might have implicitly benefited from seeing the checklists without explicitly using them, it is safe to expect that a large portion of the students did not engaged with the complementary approach. While disappointing, lack of engagement does seem to be a challenge with many educational technologies. In general, one of the most significant barriers to deploying new tools in learning environments is technology acceptance, which is defined as the willingness to use technology intended to help with tasks [84]. In the technology acceptance model in education, several factors of digital technology adoption such as attitudes towards technology, behavioural intention, perceived usefulness and ease of use should be examined [19, 75]. Winne [90] further adds that students will not engage in a new tool unless they understand its value and have the required skills to use it, so part of the challenge might that students to not realise the benefits of engaging with the approach.

Furthermore, the data showed that the system flagged less than 20% of comments. This result is promising as it suggests that we anticipated around 80% of the reviews to have included some suggestions or specific detail related to the resource under review in their feedback. However, the majority of students opted to submit the flagged comments without revision. An interesting observation is that students from INFS were more likely to revise their flagged comments. Various reasons might have contributed to this difference, such as instructions provided by educators or the different assessment requirements where INFS students had to do half the amount of reviews of the NEUR students. Further exploration, as discussed in section 5.1, is required so that we can consider the more diverse features of the textual feedback in the automatic quality control functions. Our results reiterate the findings of [42] suggesting that while considering learning theories and pedagogical approaches is important for developing educational technologies, other factors that contribute to acceptance and useability are also critical. Tsai [86] also underlined that the success of e-learning is mainly dependent on student acceptance of the system and desire to utilise it.

RQ2. Impact on feedback quality. The results of the conducted between-subject experiment in Section 4.2 validated the second hypothesis (H2) that the proposed the complementary approach would enable students to provide higher quality feedback. Findings showed that students in the experiment group (CP) wrote comments almost twice as long as comments being provided by the students in the control group (NP) who did not have access to the complementary approach. While having longer comments does not formally provide any guarantees of higher quality, the work of Zong et al. [99] reports a strong association between the feedback quality and the length of the provided comments. Additionally, Zhu and Carless [96] argue that providing lengthy comments, regardless of the quality, benefits the reviewer. This benefit may be partly attributed to the fact that they have put more effort into completing a review, which again can contribute to learning and self-regulation [6]. These findings are further corroborated by Cavalcanti et al. [14], Osakwe et al. [65], who indicated that features related to length, such as the number of words per sentence and the overall number of function words, best represent feedback about the process, which is also considered the most effective feedback level [30]. It is argued that offering students a larger scenario will help them develop the self-regulation skills needed to come up with their own solutions to problems. Kovanović et al. [45] also discovered that the number of words in student online discussion transcripts was the best predictor of quality in terms of cognitive presence, which was consistent with previous research on automated essay assessments [66]. In addition, the results of the conducted study (see Fig 6b) revealed that the comments provided by students in the experiment group (CP) were three times more likely to be perceived as helpful than the ones provided by the control group (NP). Furthermore, while Cramer's V reveals only small effect sizes for feedback qualities like alignment, suggestion, and constructive language, it reveals a stronger relationship between the peer review condition and providing detailed and specific feedback. Being specific and detailed in feedback is considered a dominant feature to increase feedback implementation [63] and asserted to be perceived more valuable than generic praises or criticisms [31]. These consistently observed enhancements in the various quantitative (e.g., length) and qualitative (e.g., specificity) features of the comments from the CP group reviews suggest the complementary approach's success in assisting students to provide better and more helpful feedback.

RQ3. Impact on judgement. Our third hypothesis (H3) was that if students were more engaged with the feedback, they would be less lenient and more confident in their reviews. Part of our assumption was that more detailed comments might make students more critical, resulting in a drop in their decision ratings. Similarly, we hypothesised that providing more detailed comments would increase students' trust in their judgement, resulting in higher confidence ratings. Our research, on the other hand, contradicts this hypothesis. These results supported the commonly reported issue of the leniency bias (the inclination to offer mostly positive ratings) in self and peer assessment [25, 56, 57, 59]. de Moira et al. [21] also found that reviewers' leniency is relatively stable over time. An interesting and perhaps related observation here is that ratings from the experiment group (CP) had a larger standard deviation than those in the NP group, suggesting that students in the experiment group were more likely to provide more extreme (high or low) ratings.

5.1 Limitation and Future Work

We see three main limitations to the current study. One, the study explored the impact of providing students with self-monitoring and automatic quality control processes simultaneously as a unified model. This approach makes it infeasible to determine the impact of each of these treatments in isolation on students feedback and judgment ability. Accordingly, one interesting direction that could be followed in the future would be to extend the conducted controlled study to five experimental conditions: rubric-only, training, self-monitoring, oversight, and the approach from this study that has all combined into one. Second, other than relatedness and explicit suggestions, the automatic quality control functions do not consider other aspects of quality feedback. Accordingly, future research could address this limitation by training and fine-tuning NLP models with manually coded comments to measure additional features of reviewers feedback, such as alignment and constructive language usage. Third, the conducted study did not take into account the demographic and personal information of students as they were out of the context of the conducted study. However, as it was elaborated by [78], differences in students' features such as their language proficiency (Native speaker vs English as the second language speaker) could impact the way that students interpret and apply the provided feedback. Accordingly, an interesting future direction would be to conducting controlled experiments that takes demographic and personal information of students into account when interpreting the results.

Furthermore, some points restrict the generalisability of the presented findings. First, while much care has been taken to ensure that the control and experiment groups are comparable, there may be external factors that the authors are unaware of that may have contributed to differences in the results between the groups. A future direction can be to replicate the study as a randomised experiment in a lab setting without ethical concerns regarding disadvantaging students' learning in a particular group. Second, the experiment considered data only from two courses in specific domains. Future work aims to replicate the study with participants from over ten courses in other subject domains and pedagogical contexts that have adopted RiPPLE. Finally, while the experiment in this paper applied the proposed complementary approach in peer review of the student-generated content in a learnersourcing system, it is a context- and domain-independent approach. We hypothesise that the complementary approach can be used in other systems/platforms that support peer review. Future works can investigate this potential in other fields and other technology-enhanced learning platforms.

ACKNOWLEDGMENTS

This research was supported partially by the Australian Government through the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) project number IC200100022.

REFERENCES

[1] Solmaz Abdi, Hassan Khosravi, and Shazia Sadiq. 2020. Modelling learners in crowdsourcing educational systems. In *International Conference on Artificial Intelligence in Education*. Springer, 3–9.

[2] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Ali Darvishi. 2021. Open Learner Models for Multi-Activity Educational Systems. In *International Conference on Artificial Intelligence in Education*. Springer, 11–17.

[3] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Gianluca Demartini. 2021. Evaluating the Quality of Learning Resources: A Learnersourcing Approach. *IEEE Transactions on Learning Technologies* 14, 1 (2021), 81–92. https://doi.org/10.1109/TLT.2021.3058644

[4] Vincent Aleven, Elizabeth A McLaughlin, R Amos Glenn, and Kenneth R Koedinger. 2016. Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction* (2016), 522–560.

[5] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[6] Martine Baars, Lisette Wijnia, Anique de Bruin, and Fred Paas. 2020. The relation between student's effort and monitoring judgments during learning: a meta-analysis. *Educational Psychology Review* (2020), 1–24.

[7] Simon P Bates, Ross K Galloway, Jonathan Riise, and Danny Homer. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics-Physics Education Research* 10, 2 (2014), 020105.

[8] David Boud, Ruth Cohen, and Jane Sampson. 2014. *Peer learning in higher education: Learning from and with each other*. Routledge.

[9] Susan Bull. 2020. There Are Open Learner Models About! *IEEE Transactions on Learning Technologies* (2020).

[10] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] David Carless. 2009. Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education* 34, 1 (2009), 79–89.

[12] David Carless. 2020. From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education* (2020).

[13] David Carless, Kennedy Kam Ho Chan, Jessica To, Margaret Lo, and Elizabeth Barrett. 2018. Developing students' capacities for evaluative judgement through analysing exemplars. In *Developing evaluative judgement in Higher Education*. Routledge, 108–116.

[14] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 428–437.

[15] Yea-huey Chang et al. 2016. Two decades of research in L2 peer review. *Journal of Writing Research* 8, 1 (2016).

[16] Kwangsu Cho and Charles MacArthur. 2011. Learning by reviewing. *Journal of Educational Psychology* 103, 1 (2011), 73.

[17] Ali Darvishi, Hassan Khosravi, and Shazia Sadiq. 2020. Utilising Learnersourcing to Inform Design Loop Adaptivity. In *European Conference on Technology Enhanced Learning*. Springer, 332–346.

[18] Ali Darvishi, Hassan Khosravi, and Shazia Sadiq. 2021. Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 139–150.

[19] Ali Darvishi, Hassan Khosravi, Shazia Sadiq, and Barbara Weber. 2021. Neurophysiological Measurements in Higher Education: A Systematic Literature Review. *International Journal of Artificial Intelligence in Education* (2021), 1–41.

[20] Luca De Alfaro and Michael Shavlovsky. 2014. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education*. 415–420.

[21] Anne Pinot de Moira, Claire Massey, Jo-Anne Baird, and Marie Morrissy. 2002. Marking consistency over time. *Research in Education* 67, 1 (2002), 79–87.

[22] Paul Denny, Andrew Luxton-Reilly, and Beth Simon. 2009. Quality of student contributed questions using PeerWise. In *Proceedings of the Eleventh Australasian Conference on Computing Education-Volume 95*. 55–63.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[24] Erkan Er, Yannis Dimitriadis, and Dragan Gašević. 2020. A collaborative learning approach to dialogic peer feedback: a theoretical framework. *Assessment & Evaluation in Higher Education* (2020), 1–15.

[25] Aslihan Erman Aslanoglu, Ismail Karakaya, and Mehmet Sata. 2020. Evaluation of University Students' Rating Behaviors in Self and Peer Rating Process via Many Facet Rasch Model. *Eurasian Journal of Educational Research* 89 (2020), 25–46.

[26] Kyle W Galloway and Simon Burns. 2015. Doing it for themselves: students creating a high quality peer-learning environment. *Chemistry Education Research and Practice* 16, 1 (2015), 82–92.

[27] Elena L Glassman, Aaron Lin, Carrie J Cai, and Robert C Miller. 2016. Learnersourcing personalized hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1626–1636.

[28] George Gyamfi, Barbara Hanna, and Hassan Khosravi. 2021. Supporting peer evaluation of student-generated content: a study of three approaches. *Assessment & Evaluation in Higher Education* 0, 0 (2021), 1–19. https://doi.org/10.1080/02602938.2021.2006140

[29] George Gyamfi, Barbara E Hanna, and Hassan Khosravi. 2021. The effects of rubrics on evaluative judgement: a randomised controlled experiment. *Assessment & Evaluation in Higher Education* 47, 1 (2021), 126–143.

[30] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[31] Michael Henderson, Michael Phillips, Tracii Ryan, David Boud, Phillip Dawson, Elizabeth Molloy, and Paige Mahoney. 2019. Conditions that enable effective feedback. *Higher Education Research & Development* 38, 7 (2019), 1401–1416.

[32] Ken Hyland. 2019. *Second language writing*. Cambridge university press.

[33] Renée S Jansen, Anouschka Van Leeuwen, Jeroen Janssen, Suzanne Jak, and Liesbeth Kester. 2019. Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review* 28 (2019), 100292.

[34] Lasse X Jensen, Margaret Bearman, and David Boud. 2021. Understanding feedback in online learning–A critical review and metaphor analysis. *Computers & Education* (2021), 104271.

[35] Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. *arXiv preprint arXiv:2110.03895* (2021).

[36] David A Joyner. 2017. Scaling expert feedback: Two case studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 71–80.

[37] Douglas F Kauffman, Ruomeng Zhao, and Ya-Shu Yang. 2011. Effects of online note taking formats and self-monitoring prompts on learning from online text: Using technology to enhance self-regulated learning. *Contemporary Educational Psychology* 36, 4 (2011), 313–322.

[38] Hassan Khosravi, Kendra Cooper, and Kirsty Kitto. 2017. RiPLE: Recommendation in Peer-Learning Environments Based on Knowledge Gaps and Interests. *JEDM| Journal of Educational Data Mining* 9, 1 (2017), 42–67.

[39] Hassan Khosravi, Gianluca Demartini, Shazia Sadiq, and Dragan Gasevic. 2021. Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 32–42.

[40] Hassan Khosravi, George Gyamfi, Barbara E. Hanna, Jason Lodge, and Solmaz Abdi. 2021. Bridging the Gap Between Theory and Empirical Research in Evaluative Judgment. *Journal of Learning Analytics* 8, 3 (2021), 117–132.

[41] Hassan Khosravi, Kirsty Kitto, and Williams Joseph. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics* 6, 3 (2019), 91–105.

[42] Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Development and adoption of an adaptive learning system: Reflections and lessons learned. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 58–64.

[43] Juho Kim. 2015. *Learnersourcing: improving learning with collective learner activity*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[44] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 4017–4026.

[45] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 15–24.

[46] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4627–4639.

[47] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 6 (2013), 1–31.

[48] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 75–84.

[49] Hatim Lahza, Hassan Khosravi, Gianluca Demartini, and Dragan Gasevic. 2022. Effects of Technological Interventions for Self-regulation: A Control Experiment in Learnersourcing. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 542–548.

[50] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[51] Hongli Li, Yao Xiong, Charles Vincent Hunter, Xiuyan Guo, and Rurik Tywoniw. 2020. Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education* 45, 2 (2020), 193–211.

[52] Hsien-Chin Liou and Zhong-Yan Peng. 2009. Training effects on computer-mediated peer review. *System* 37, 3 (2009), 514–525.

[53] Ngar-Fun Liu and David Carless. 2006. Peer feedback: the learning element of peer assessment. *Teaching in Higher education* 11, 3 (2006), 279–290.

[54] Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of second language writing* 18, 1 (2009), 30–43.

[55] Mario Manso-Vázquez and Martin Llamas-Nistal. 2015. A monitoring system to ease self-regulated learning processes. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 10, 2 (2015), 52–59.

[56] Sumie Matsuno. 2021. The Effects of Informing the Quality of Students' Previous Peer Assessment. In *JACET International Convention Selected Papers*, Vol. 1. The Japan Association of College English Teachers, 174.

[57] Amy McMillan, Pol Solanelles, and Bryan Rogers. 2021. Bias in student evaluations: Are my peers out to get me? *Studies in Educational Evaluation* 70 (2021), 101032.

[58] Cristina Mercader, Georgeta Ion, and Anna Díaz-Vicario. 2020. Factors influencing students' peer feedback uptake: instructional design matters. *Assessment & Evaluation in Higher Education* (2020), 1–12.

[59] Rafael Molina-Carmona, Rosana Satorre-Cuerda, PATRICIA Compañ-Rosique, and Faraón Llorens-Largo. 2018. Metrics for estimating validity, reliability and bias in peer assessment. *International Journal of Engineering Education* 34, 3 (2018).

[60] T. Moon. 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* 13 (1996), 47–60.

[61] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 588–593.

[62] Sapna Negi, Kartik Asooja, Shubham Mehrotra, and Paul Buitelaar. 2016. A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 170–178.

[63] Melissa M Nelson and Christian D Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional science* 37, 4 (2009), 375–401.

[64] David Nicol, Avril Thomson, and Caroline Breslin. 2014. Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education* 39, 1 (2014), 102–122. https://doi.org/10.1080/02602938.2013.795518

[65] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gaševic, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2021. Towards automated content analysis of feedback: A multi-language study. In *Proceedings of the 14th International Conference on Educational Data Mining*.

[66] Ellis B Page and Nancy S Petersen. 1995. The computer moves into essay grading: Updating the ancient test. *Phi delta kappan* 76, 7 (1995), 561.

[67] Jennifer Palisse, Deborah Martina King, and Mark MacLean. 2021. Comparative judgement and the hierarchy of students' choice criteria. *International Journal of Mathematical Education in Science and Technology* (2021), 1–21.

[68] Ernesto Panadero, Jesús Alonso Tapia, and Juan Antonio Huertas. 2012. Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and individual differences* 22, 6 (2012), 806–813.

[69] Dwayne E Paré and Steve Joordens. 2008. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning* 24, 6 (2008), 526–540.

[70] Melissa M Patchan and Christian D Schunn. 2015. Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science* 43, 5 (2015), 591–614.

[71] Helen Purchase and John Hamer. 2018. Peer-review in practice: eight years of Aropä. *Assessment & Evaluation in Higher Education* 43, 7 (2018), 1146–1165.

[72] Lakshmi Ramachandran, Edward F Gehringer, and Ravi K Yadav. 2017. Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education* 27, 3 (2017), 534–581.

[73] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[74] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[75] R. Scherer, F. Siddiq, and J. Tondeur. 2019. The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education* 128 (2019), 13–35.

[76] Bernhard Schmitz and Franziska Perels. 2011. Self-monitoring of self-regulation during math homework behaviour using standardized diaries. *Metacognition and Learning* 6, 3 (2011), 255–273.

[77] A Shibani. 2019. *Augmenting Pedagogic Writing Practice with Contextualizable Learning Analytics*. Ph. D. Dissertation.

[78] Antonette Shibani, Simon Knight, and Simon Buckingham Shum. 2020. Educator perspectives on learning analytics in classroom practice. *The Internet and Higher Education* 46 (2020), 100730.

[79] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. 2016. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. 179–196.

[80] Victor Shnayder and David C Parkes. 2016. Practical peer prediction for peer assessment. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[81] Nancy Sommers. 1982. Responding to student writing. *College composition and communication* 33, 2 (1982), 148–156.

[82] Sean Tackett, Mark Raymond, Rishi Desai, Steven A Haist, Amy Morales, Shiv Gaglani, and Stephen G Clyman. 2018. Crowdsourcing for assessment items to support adaptive learning. *Medical teacher* 40, 8 (2018), 838–841.

[83] Joanna Tai, Rola Ajjawi, David Boud, Phillip Dawson, and Ernesto Panadero. 2018. Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education* 76, 3 (2018), 467–481.

[84] Timothy Teo. 2011. *Technology acceptance in education*. Springer Science & Business Media.

[85] Keith Topping. 2003. Self and peer assessment in school and university: Reliability, validity and utility. In *Optimising new modes of assessment: In search of qualities and standards*. Springer, 55–87.

[86] Yea-Ru Tsai. 2015. Applying the Technology Acceptance Model (TAM) to explore the effects of a Course Management System (CMS)-Assisted EFL writing instruction. *Calico Journal* 32, 1 (2015), 153–171.

[87] Jason L Walsh, Benjamin HL Harris, Paul Denny, and Phil Smith. 2018. Formative student-authored question bank: perceptions, question quality and association with summative performance. *Postgraduate medical journal* 94, 1108 (2018), 97–103.

[88] Wanyuan Wang, Bo An, and Yichuan Jiang. 2020. Optimal Spot-Checking for Improving the Evaluation Quality of Crowdsourcing: Application to Peer Grading Systems. *IEEE Transactions on Computational Social Systems* 7, 4 (2020), 940–955.

[89] David Kofoed Wind, Rasmus Malthe Jørgensen, and Simon Lind Hansen. 2018. Peer Feedback with Peergrade. In *ICEL 2018 13th International Conference on e-Learning*. Academic Conferences and publishing limited, 184.

[90] Philip H Winne. 2006. How software technologies can improve research on learning and bolster school reform. *Educational psychologist* 41, 1 (2006), 5–17.

[91] James R Wright, Chris Thornton, and Kevin Leyton-Brown. 2015. Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 96–101.

[92] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[93] Wenting Xiong, D Litmaan, and Christian Schunn. 2012. Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research* 4, 2 (2012), 155–176.

[94] David Scott Yeager, Valerie Purdie-Vaughns, Julio Garcia, Nancy Apfel, Patti Brzustoski, Allison Master, William T Hessert, Matthew E Williams, and Geoffrey L Cohen. 2014. Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General* 143, 2 (2014), 804.

[95] Meina Zhu and Curtis J Bonk. 2019. Designing MOOCs to Facilitate Participant Self-Monitoring for Self-Directed Learning. *Online Learning* 23, 4 (2019), 106–134.

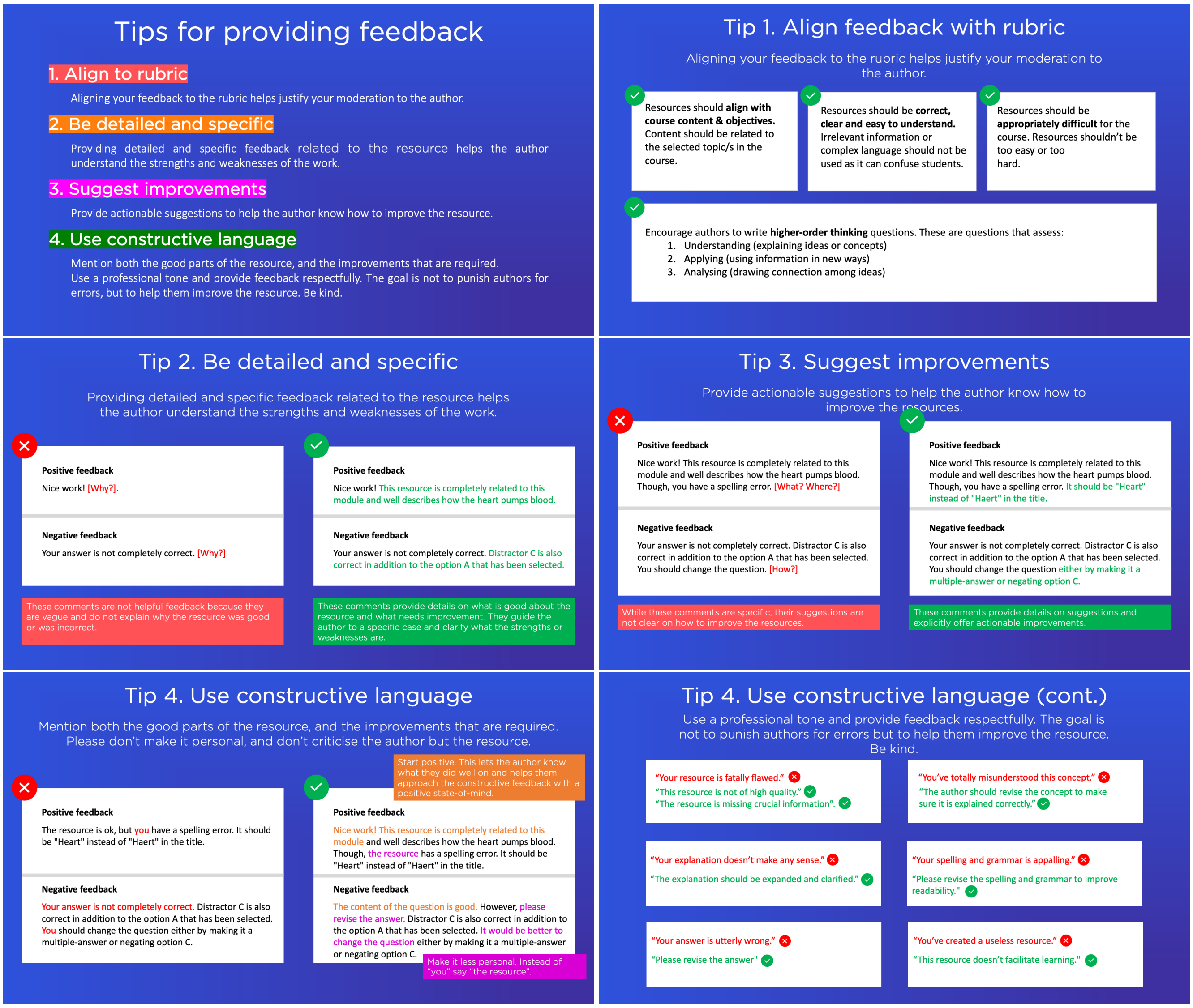
[96] Qiyun Zhu and David Carless. 2018. Dialogue within peer feedback processes: Clarification and negotiation of meaning. *Higher Education Research & Development* 37, 4 (2018), 883–897.

[97] Barry J Zimmerman, Sebastian Bonner, and Robert Kovach. 1996. *Developing self-regulated learners: Beyond achievement to self-efficacy.* American Psychological Association.

[98] Barry J Zimmerman and Andrew S Paulsen. 1995. Self-monitoring during collegiate studying: An invaluable tool for academic self-regulation. *New directions for teaching and learning* 1995, 63 (1995), 13–27.

[99] Zheng Zong, Christian D Schunn, and Yanqing Wang. 2021. What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior* (2021), 106924.

A TRAINING MATERIALS



Tips for providing feedback

1. Align to rubric

Aligning your feedback to the rubric helps justify your moderation to the author.

2. Be detailed and specific

Providing detailed and specific feedback related to the resource helps the author understand the strengths and weaknesses of the work.

3. Suggest improvements

Provide actionable suggestions to help the author know how to improve the resource.

4. Use constructive language

Mention both the good parts of the resource, and the improvements that are required.
Use a professional tone and provide feedback respectfully. The goal is not to punish authors for errors, but to help them improve the resource. Be kind.

Tip 1. Align feedback with rubric

Aligning your feedback to the rubric helps justify your moderation to the author.

✓ Resources should **align with course content & objectives.** Content should be related to the selected topic/s in the course.

✓ Resources should be **correct, clear and easy to understand.** Irrelevant information or complex language should not be used as it can confuse students.

✓ Resources should be **appropriately difficult** for the course. Resources shouldn't be too easy or too hard.

✓ Encourage authors to write **higher-order thinking** questions. These are questions that assess:

1. Understanding (explaining ideas or concepts)
2. Applying (using information in new ways)
3. Analysing (drawing connection among ideas)

Tip 2. Be detailed and specific

Providing detailed and specific feedback related to the resource helps the author understand the strengths and weaknesses of the work.

✗ **Positive feedback**

Nice work! [Why?].

Negative feedback

Your answer is not completely correct. [Why?]

These comments are not helpful feedback because they are vague and do not explain why the resource was good or was incorrect.

✓ **Positive feedback**

Nice work! This resource is completely related to this module and well describes how the heart pumps blood.

Negative feedback

Your answer is not completely correct. Distractor C is also correct in addition to the option A that has been selected.

These comments provide details on what is good about the resource and what needs improvement. They guide the author to a specific case and clarify what the strengths or weaknesses are.

Tip 3. Suggest improvements

Provide actionable suggestions to help the author know how to improve the resources.

✗ **Positive feedback**

Nice work! This resource is completely related to this module and well describes how the heart pumps blood. Though, you have a spelling error. [What? Where?]

Negative feedback

Your answer is not completely correct. Distractor C is also correct in addition to the option A that has been selected. You should change the question. [How?]

While these comments are specific, their suggestions are not clear on how to improve the resources.

✓ **Positive feedback**

Nice work! This resource is completely related to this module and well describes how the heart pumps blood. Though, you have a spelling error. It should be "Heart" instead of "Haert" in the title.

Negative feedback

Your answer is not completely correct. Distractor C is also correct in addition to the option A that has been selected. You should change the question either by making it a multiple-answer or negating option C.

These comments provide details on suggestions and explicitly offer actionable improvements.

Tip 4. Use constructive language

Mention both the good parts of the resource, and the improvements that are required. Please don't make it personal, and don't criticise the author but the resource.

✗ **Positive feedback**

The resource is ok, but you have a spelling error. It should be "Heart" instead of "Haert" in the title.

Negative feedback

Your answer is not completely correct. Distractor C is also correct in addition to the option A that has been selected. You should change the question either by making it a multiple-answer or negating option C.

Start positive. This lets the author know what they did well on and helps them approach the constructive feedback with a positive state-of-mind.

✓ **Positive feedback**

Nice work! This resource is completely related to this module and well describes how the heart pumps blood. Though, the resource has a spelling error. It should be "Heart" instead of "Haert" in the title.

Negative feedback

The content of the question is good. However, please revise the answer. Distractor C is also correct in addition to the option A that has been selected. It would be better to change the question either by making it a multiple-answer or negating option C.

Make it less personal. Instead of "you" say "the resource".

Tip 4. Use constructive language (cont.)

Use a professional tone and provide feedback respectfully. The goal is not to punish authors for errors but to help them improve the resource. Be kind.

"Your resource is fatally flawed." ✗
"This resource is not of high quality." ✓
"The resource is missing crucial information". ✓

"You've totally misunderstood this concept." ✗
"The author should revise the concept to make sure it is explained correctly." ✓

"Your explanation doesn't make any sense." ✗
"The explanation should be expanded and clarified." ✓

"Your spelling and grammar is appalling." ✗
"Please revise the spelling and grammar to improve readability." ✓

"Your answer is utterly wrong." ✗
"Please revise the answer" ✓

"You've created a useless resource." ✗
"This resource doesn't facilitate learning." ✓

Figure 9: Tips for providing feedback