Received: 5 December 2021 | Accepted: 23 April 2022

DOI: 10.1111/bjet.13233

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Incorporating AI and learning analytics to build trustworthy peer assessment systems

Ali Darvishi[1] | Hassan Khosravi[1] | Shazia Sadiq[1] | Dragan Gašević[2]

[1]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Queensland, Australia

[2]Faculty of Information Technology, Monash University, Melbourne, Victoria, Australia

Correspondence
Ali Darvishi and Hassan Khosravi, School of Information Technology and Electrical Engineering and Data Science and Learning Analytics Institute for Teaching and Learning Innovation, The University of Queensland, Brisbane, QLD 4072, Australia.
Email: a.darvishi@uq.edu.au and h.khosravi@uq.edu.au

Funding information
This article is funded by The University of Queensland.

Abstract

Peer assessment has been recognised as a sustainable and scalable assessment method that promotes higher-order learning and provides students with fast and detailed feedback on their work. Despite these benefits, some common concerns and criticisms are associated with the use of peer assessments (eg, scarcity of high-quality feedback from peer student-assessors and lack of accuracy in assigning a grade to the assessee) that raise questions about their trustworthiness. Consequently, many instructors and educational institutions have been anxious about incorporating peer assessment into their teaching. This paper aims to contribute to the growing literature on how AI and learning analytics may be incorporated to address some of the common concerns associated with peer assessment systems, which in turn can increase their trustworthiness and adoption. In particular, we present and evaluate our AI-assisted and analytical approaches that aim to (1) offer guidelines and assistance to student-assessors during individual reviews to provide better feedback, (2) integrate probabilistic and text analysis inference models to improve the accuracy of the assigned grades, (3) develop feedback on reviews strategies that enable peer assessors to review the work

Ali Darvishi and Hassan Khosravi contributed equally to this study.

of each other, and (4) employ a spot-checking mechanism to assist instructors in optimally overseeing the peer assessment process.

KEYWORDS
human centred AI, learning analytics, peer assessment

Practitioner notes

What is already known about this topic

- Engaging students in peer assessment has been demonstrated to have various benefits. However, there are some common concerns associated with employing peer assessment that raise questions about their trustworthiness as an assessment item.

What this paper adds

- Methods and processes on how AI and learning analytics may be incorporated to address some of the common concerns associated with peer assessment systems, which in turn, can increase their trustworthiness and adoption.

Implications for practice

- Presentation of a systematic approach for development, deployment and evaluation of AI and analytics approaches in peer assessment systems.

INTRODUCTION

Peer assessment can formally be defined as “an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners” (Topping, 2009, p. 20). Peer assessment can be incorporated under various settings such as online or in-person (Yu & Wu, 2011), pairs or groups (Topping, 2010), open or blinded (Shoham & Pitman, 2021) on a wide range of learning activities including oral presentations (Nejad & Mahfoodh, 2019), essays (Huisman et al., 2018), portfolios (Yang et al., 2016), projects (Lin, 2018) and student-generated content (Darvishi et al., 2021).

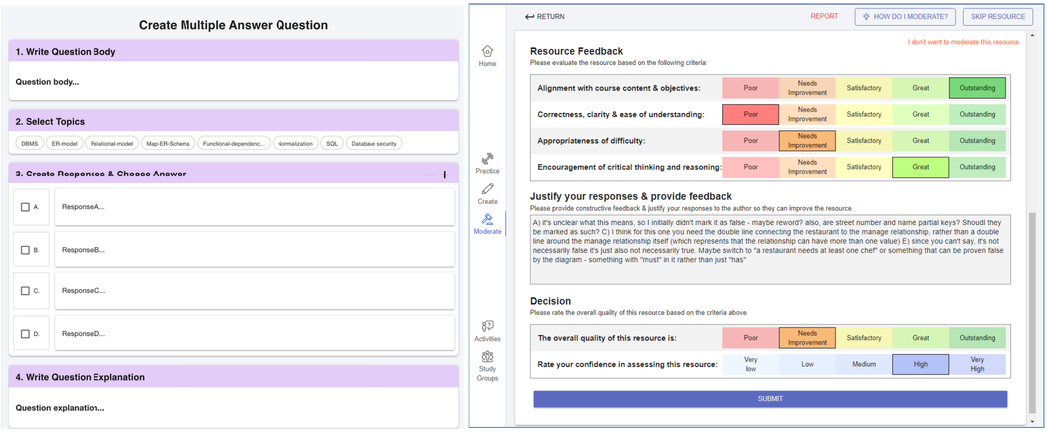
Engaging students in peer assessment has been demonstrated to have various benefits for the assessors, assessees and instructors. Assessors gain the opportunity to improve their comprehension of the content (Li et al., 2010), develop a greater sense of accountability (Kao, 2013) and evaluative judgement (Nicol et al., 2014; Tai et al., 2018; Khosravi et al., 2021; Gyamfi et al., 2021), improve their writing (Polisda, 2017), and gain experience in providing constructive feedback (Lundstrom & Baker, 2009). Assessees gain the opportunity to receive more immediate and individualised feedback (Kulkarni et al., 2015) from peers with diverse perspectives (Cho & MacArthur, 2011; Patchan & Schunn, 2015), which are perceived as less authoritative and more open to a reciprocal exchange of views and negotiation (Topping, 2009). Finally, instructors' workload related to student marking can be reduced, which creates opportunities to increase enrollment (Joyner, 2017) or for them to reinvest their time more optimally towards enhancing student learning. Additionally, the data generated from students' engagement with the peer assessment process may be utilised by learning analytics tools and learning analytics dashboards (Matcha et al., 2019) to complement the clickstream data captured by learning management systems to enable instructors to gain insights into students learning process. Due to these and other benefits, many peer assessment systems such a PeerScholar (Paré & Joordens, 2008), Mechanical TA (Wright et al., 2015), Aropä (Purchase & Hamer, 2018), CrowdGrader (De Alfaro & Shavlovsky, 2014) and Peergrade (Wind et al., 2018) that facilitate peer review activities have been developed.

Despite these benefits, there are some common concerns and criticisms associated with the use of peer assessment systems. Some of the main concerns target the nature of having students as experts in training and are rooted in the logical argument that they may lack the required knowledge, incentive or ability to evaluate their peers' work accurately and to provide effective feedback (Carless, 2009; Patchan et al., 2018; Sridharan et al., 2019). In peer assessment systems where multiple reviewers review the same task, there are concerns associated with how to accurately assign a final grade as basic aggregation approaches such as mean and median that assume equal weight for all reviewers have been shown to be ineffective (Abdi et al., 2020; Darvishi et al., 2021; Topping, 2009). Additionally, many of the peer review systems do not formally provide the opportunity for the assessee and other assessors to raise concerns and provide feedback on the reviews. In those that do, it is challenging to develop processes and analytics that help instructors and the system to identify the main concerns raised by the students. Finally, as a follow up to the previous point, instructors may not be able to effectively oversee the review process and to identify cases where peer assessors might have made a poor judgement (Wang et al., 2018) without being overwhelmed by additional work. All of these aspects lead to reducing trust defined as “a firm belief in the competence of an entity to act reliably” (Robinson, 1996, p. 3), in peer assessment systems. Consequently, many instructors and educational institutions have been hesitant about incorporating them into their teaching (Liu & Carless, 2006).

In this paper, we explore the potential of using AI and learning analytics to address some of the common concerns discussed above as a potential approach to increase the trustworthiness of peer assessment systems. We focus on AI-assisted and analytical approaches related to four processes carried out in peer assessments systems: (1) *Individual reviews*, where assessors individually complete the assessment task, (2) *Assigning grades*, where the system automatically aggregates assessors' decisions to assign a final grade, (3) *Feedback on reviews*, where the assessee and assessors provide feedback on peer reviews, discuss the outcome, and raise concerns, and (4) *instructor oversight*, where the instructor reviews the work of assessors and oversees the peer assessment process. We implement and evaluate our proposed approaches using a learning tool called RiPPLE that supports peer assessment. Our evaluation of the proposed approaches for each of the four processes is guided by a set of research questions, where a range of techniques such as conducting controlled experiments, thematic analysis of student comments and descriptive analysis of data collected by RiPPLE are used for answering these research questions. We conclude by examining benefits, implications and potential challenges and points for consideration in relation to incorporating AI and learning analytics for building trustworthy peer assessment systems.

THE RiPPLE PLATFORM

RiPPLE is an adaptive educational system that dynamically modifies the level or form of instruction based on individual student skills or preferences in order to deliver a personalised learning experience (Khosravi et al., 2019). Adaptive educational systems require an extensive repository of learning resources, often developed by domain experts, to provide customised learning for students with varying knowledge levels. RiPPLE, on the other hand, relies heavily on the learnersourcing approach, which involves students in the creation and evaluation of a variety of learning resources (Khosravi et al., 2021; Abdi et al., 2021). RiPPLE allows both students and instructors to generate learning resources. Users may



Question body...

□ A. ResponseA...

□ B. ResponseB...

□ C. ResponseC...

□ D. ResponseD...

Question explanation...

(a) Resource creation interface

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.

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

(b) Peer assessment interface

FIGURE 1 Interfaces of the three main processes in RiPPLE

create multiple-choice questions, multiple-answer questions, working examples, and open-ended notes. The interface for creating multiple answer questions is depicted in Figure 1a.

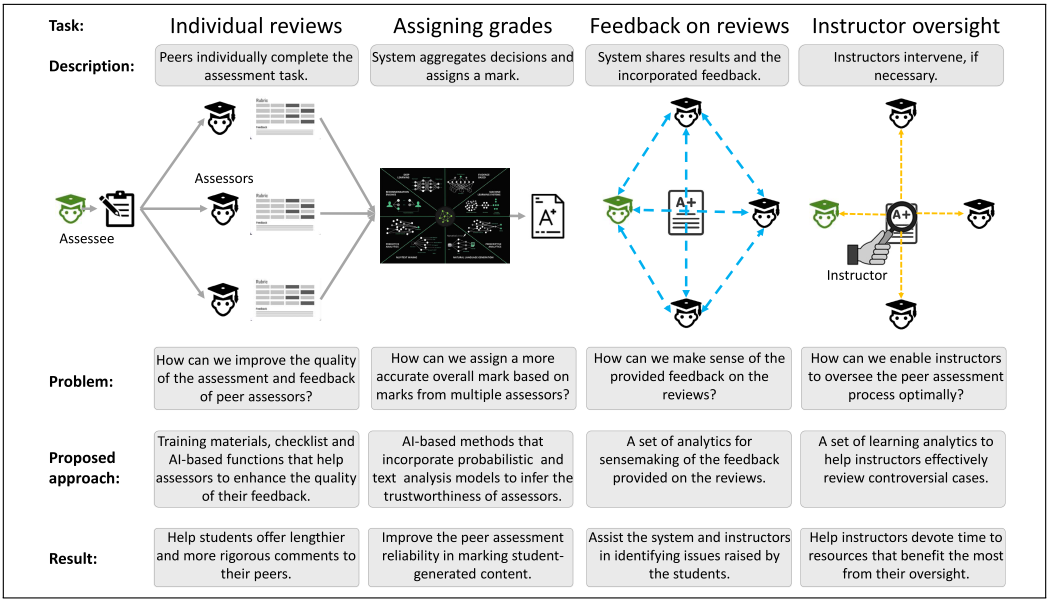
An instructor-created resource is directly uploaded to the repository of learning materials, but a student-created resource must go through a peer assessment procedure. Users can create different types of resources, including multiple-choice questions, multi-answer questions, worked examples as well as open-ended notes. Peer assessors are requested to evaluate the quality of student-generated content using a multiple-criteria rubric that instructors can customise for each resource type. For example, Figure 1b shows the assessment interface peers use to evaluate a multiple-choice question by considering alignment, accuracy, difficulty level and critical thinking encouragement. Before submitting their assessment, assessors are required to support their judgement and offer comments to the author. Finally, assessors express their final decision and level of confidence in their rating. RiPPLE determines whether it is ready to make a final decision on the quality of a resource under assessment based on the number of peer reviews received, the reliability of the assessors and the degree of agreement between the received assessments.

BUILDING TRUSTWORTHY PEER ASSESSMENT SYSTEMS

This section presents AI-assisted and analytics-driven approaches for addressing concerns with the individual reviews, assigning grades, feedback on reviews and instructor oversight processes in peer assessment tools.[1] For each of the processes, we discuss common concerns, propose solutions and present a corresponding implementation in RiPPLE, and share results via evaluating the approach using guiding research questions.[2] Figure 2 provides a graphical summary of the content presented in this section.

Individual reviews

An important step in the peer assessment process is when each assessor individually completes the assessment task. Previous work has shown that assessors may lack the incentive to do a rigorous job (Patchan et al., 2018) or lack the ability to provide effective feedback



Task:	Individual reviews	Assigning grades	Feedback on reviews	Instructor oversight
Description:	Peers individually complete the assessment task.	System aggregates decisions and assigns a mark.	System shares results and the incorporated feedback.	Instructors intervene, if necessary.
Problem:	How can we improve the quality of the assessment and feedback of peer assessors?	How can we assign a more accurate overall mark based on marks from multiple assessors?	How can we make sense of the provided feedback on the reviews?	How can we enable instructors to oversee the peer assessment process optimally?
Proposed approach:	Training materials, checklist and AI-based functions that help assessors to enhance the quality of their feedback.	AI-based methods that incorporate probabilistic and text analysis models to infer the trustworthiness of assessors.	A set of analytics for sensemaking of the feedback provided on the reviews.	A set of learning analytics to help instructors effectively review controversial cases.
Result:	Help students offer lengthier and more rigorous comments to their peers.	Improve the peer assessment reliability in marking student-generated content.	Assist the system and instructors in identifying issues raised by the students.	Help instructors devote time to resources that benefit the most from their oversight.

FIGURE 2 A graphical summary of the four peer assessment processes considered, common problems associated with them, our proposed approaches and results

(Carless & Boud, 2018). Students' failure to provide high-quality feedback is one of the main points of critique that undermines the trustworthiness of using peer review as an assessment tool, which leads to substantial negative consequences such as lowering standards (Yeager et al., 2014), reducing trust in the outcome (Carless, 2009), and making reviewees less likely to revise their work (Sommers, 1982). There are multiple explanations for why students may fail to provide effective feedback. 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 of peer review (Li et al., 2020). 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 that they follow the 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 (Jansen et al., 2019; Lahza et al., 2022). A third explanation is that students may lack the required incentive to provide feedback diligently (Shnayder et al., 2016). 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 (Topping, 2003).

Approach

To help address these concerns, we draw on insights from feedback literacy, self-regulation and natural language processing research to develop a complementary AI-assisted peer-review approach that incorporates training, self-monitoring, and text quality control techniques for students to consider when writing their reviews. The training material employs higher education research on feedback (Carless, 2020) to develop a set of tips with examples on writing effective feedback focusing on (1) following the criteria (alignment with rubric),

(2) being explicit and thorough (specificity), (3) offering suggestions for improvement, and (4) using constructive language. For self-monitoring, we incorporate a self-monitoring checklist, as one of the well-studied strategies from the self-regulation literature (Manso-Vázquez & Llamas-Nistal, 2015; Zimmerman et al., 1996) to encourage students to use the training material's recommendations and assist them in monitoring whether or not they have incorporated the recommendations in their feedback. Finally, inspired and informed by the success of using NLP methods in improving the quality of user reviews (Devlin et al., 2018; Napoles et al., 2015; Negi et al., 2016; Ramachandran et al., 2017; Reimers & Gurevych, 2019), we developed three quality control functions that automatically analyse the quality of provided comments and encourage students to improve their textual feedback. The first function employs the GLEU (Google BiLingual Evaluation Understudy) to calculate the similarity between the current submitted text and the previous comments of the student moderator using n-grams (Napoles et al., 2015), reducing the possibility of gaming the system by submitting the same general comment multiple times. The second function uses a rule-based approach proposed by Negi et al. (2016) to determine whether a suggestion has been made in the submitted comments. Finally, the third function measures the relatedness of the provided textual feedback to the resource context using SBERT (Reimers & Gurevych, 2019) as the encoder function to calculate the cosine similarity score. SBERT is based on a neural language model called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), pre-trained on a large language corpus to encode sentences in such a way that similar sentences are close to each other in the embedding space. Figure 3 depicts an overview of RiPPLE's interfaces for self-regulation checklist, and AI-regulation prompts.

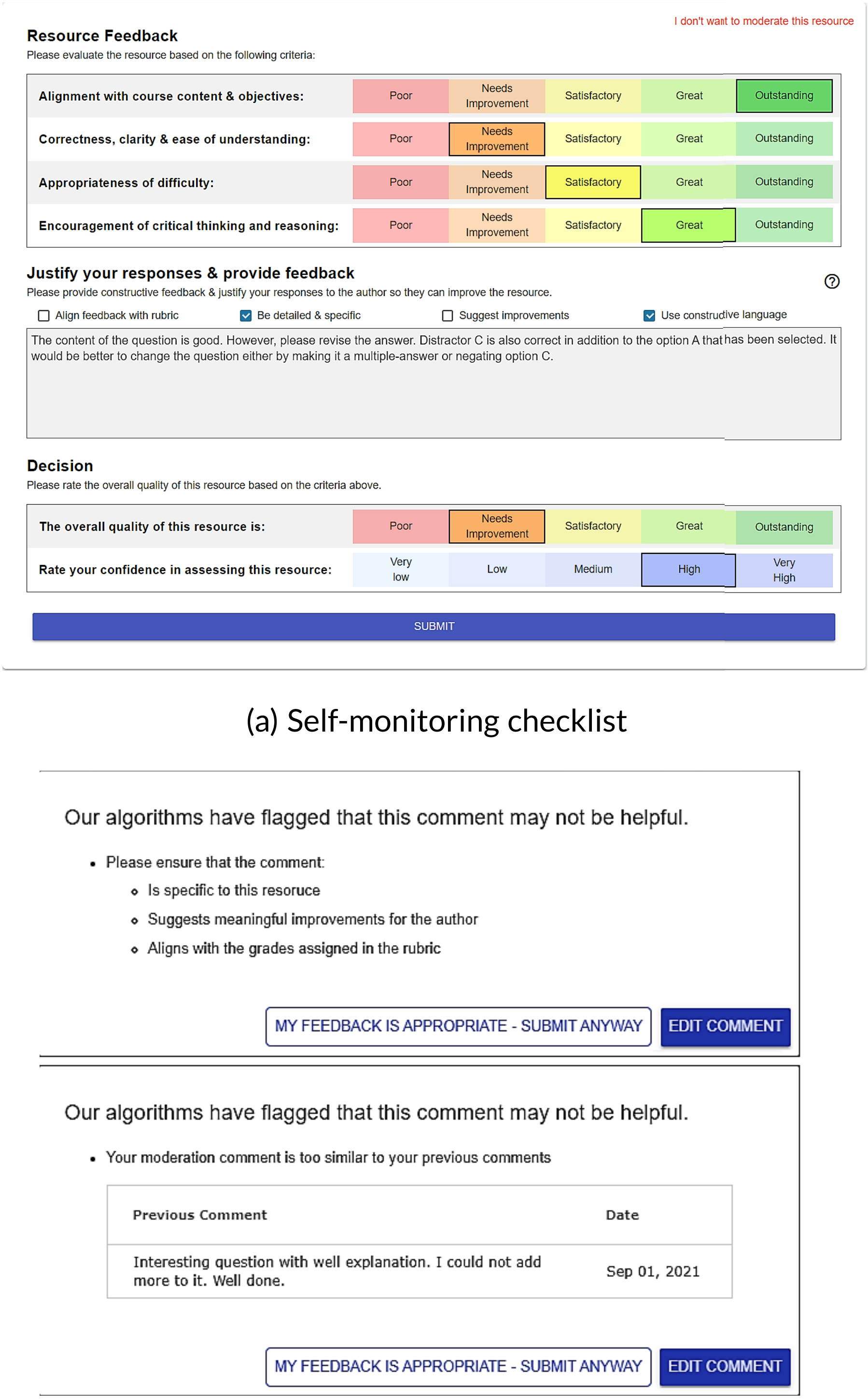
Evaluation

Our evaluation of the proposed approach was guided by the following research questions.

> **RQ1-1:** To what extent do students engage with the AI-assisted complementary approach?
>
> **RQ1-2:** What is the impact of the AI-assisted complementary approach on peer feedback?

To answer these questions, we conducted a between-subjects field experiment with $n = 374$ consented participants from two undergraduate courses, namely The Brain and Behavioural Sciences (NEUR, $n = 234$) and Introduction to Information Systems (INFS, $n = 140$). Table A1 in Appendix A provides an overview of the control and experiment groups in terms of the number of students (#Students), resources (#Resources), and peer reviews (#Reviews) in each course. The *control group* (n = 187) undertook peer reviews using the standard rubric-only interface, and the *experiment group* (n = 187) used the complemented peer review interface with the self-regulation checklist and AI-generated quality control prompts. Key highlights below and respond to the research questions below.

Response to RQ1-1: Engagement with the intervention. According to system statistics, 49.2% (ie, 92 out of 187) of students across the experiment group accessed the training material, and 35.9% (ie, 67 out of 187 students) used the self-regulation checklist at least once. According to these data, the training and self-regulation practices benefited almost half of the students directly. Some other students may have also benefited implicitly from reading the checklist without explicitly selecting the check boxes. The collected data also show that 18.2% (ie, 282 out of 1551) of comments were flagged when submitted. However, only 35.5% (ie, 100 out of 282) of flagged reviews were modified and resubmitted by students,



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 construdive 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

(a) Self-monitoring checklist

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

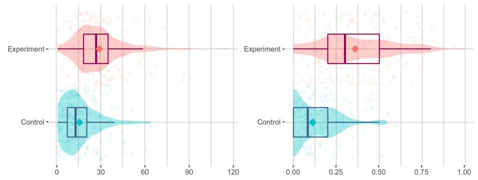
Our algorithms have flagged that this comment may not be helpful.

- 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

FIGURE 3 Complementary interface for peer review including: (a) complemented peer review interface with self-monitoring checklist, (b) automatic quality control prompts



(a) Length of Comment

(b) Like per Review Rate

Code	Group		Chi-Square Test (1, N = 317)
	Control	Experiment	
Alignment	0.62	0.79	10.4, p = .001, V = .18
Specificity	0.63	0.92	34.7, p <.05, V = .33
Suggestion	0.26	0.38	4.6, p = .031, V = .12
Constructive	0.74	0.93	18.7, p <.001, V = .24

(c) Manual coding of textual feedback quality

FIGURE 4 Comparing students' peer review behaviour in each group. (a) The average length of comments in words, (b) percentage of reviews that were perceived as helpful (the boxplots show data distributions as a five-number summary: Minimum, first quartile, median, third quartile, and maximum. The highlighted points indicate the mean. Also, the violin plots were included to display the overall distribution of the data using density curves), and (c) results of manual coding of textual feedback comparing the quality in terms of the four criteria of effective feedback for two groups

while students overlooked the tip and opted for "submit anyway" in the remaining 64.5% (ie, 182) comments.

Response to RQ1-2: Impact on feedback. Figure 4a shows how comments' length varied across experiment groups. It indicates that students in the experiment group have provided significantly longer comments (M = 29.2, SD = 16.9) than students in the control group (M = 15.8, SD = 12.2, $t(372) = 8.77$, $p < 0.001$, 95% CI [10.34, 16.32], Hedges's gs = 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, 2013]. While lengthier comments do not necessarily imply higher quality, Zong et al. (2021) identified a significant association between feedback quality and comment length. Furthermore, according to Zhu and Carless (2018), offering long comments is beneficial to the reviewer, regardless of quality. This benefit might be attributed to the fact that they worked harder to conduct the review, which, according to Baars et al. (2020), can support learning and self-regulation.

Another indication of the quality of a review and its associated feedback is the helpfulness likes and dislikes that it receives. Figure 4b 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 moderation 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), $t(372) = 13.14$, $p < 0.001$, 95% CI [0.21, 0.28], Hedges' gs = 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 peer review rate of a person from the experiment group was higher than the like peer review rate of a person from the control group.

We have further evaluated textual feedback quality by manually coding 10% of randomly selected comments from each group (ie, 163 from the control and 154 from the experiment). The codebook is built around the four criteria presented in the training materials. 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 unaware of the peer review conditions of the selected comments (ie, whether a review was in the control or the experiment group); Cohen's kappa coefficient of 0.87 indicates excellent agreement between them, with 94.6% inter-rater agreement across all codes.

The analysis indicated that 62% of peer reviews in the control group and 79% in the experiment group were 'aligned with rubrics,' as shown in Figure 3a. It also shows that the code 'detailed and specific' accounted for 63% of the comments in the control group, compared to 92% in the experiment group; 'suggested improvements' accounted for 26% in the control group, compared to 38% in the experiment group; and 74% of comments in the control group concerning 'constructive language' compared to 93% in the experiment group. Furthermore, the Chi-Square Test of Independence revealed significant relationships between the peer review conditions and comment quality across all four criteria. On the other hand, Cramer's V reveals only small effect sizes for feedback qualities such as alignment, suggestion, and constructive language, but a stronger relationship between the peer review condition and providing detailed and specific feedback, which is a dominant feature for advancing feedback execution and more practical than generic compliments or complaints (Henderson et al., 2019; Nelson & Schunn, 2009).

In summary, according to the findings, providing students with training, self-monitoring, and quality control features can help them offer lengthier and more useful comments to other peers.

Assigning grades

In cases, where peer assessment is used for summative assessment, part of the challenge is assigning a grade to the assessee (Sridharan et al., 2019). Given that judgements of students, as experts-in-training, cannot wholly be relied upon (Darvishi et al., 2021), a redundancy-based method is widely employed where the same assessment task is given to multiple students. However, this approach raises the issue of "truth inference"– How can we efficiently integrate multiple people's decisions towards an accurate final decision in the absence of ground truth? Many peer assessment systems employ summary statistics such as mean and median (eg, Paré & Joordens, 2008; Purchase & Hamer, 2018; Wind et al., 2018; Wright et al., 2015). However, these basic aggregation approaches assume equal weight for all reviewers who contributed to a peer assessment task, regardless of their abilities or interests. These basic approaches have been shown to be ineffective and sometimes lead to a lack of differentiation between high and low-quality submissions (Abdi et al., 2020; Darvishi et al., 2021; Topping, 2009).

Approach

To address this challenge, one possibility is to incorporate more advanced models with the aim of inferring the reliability of each assessor (Tao et al., 2018). Below, we present four models that can be grouped into two categories: Probabilistic models that aim to infer the trustworthiness of an assessor based on the grades they have provided and text analysis models that aim to infer the trustworthiness of an assessor based on the feedback they have given. The latter approach is employable in peer assessment tasks where assessors are expected to leave comprehensive feedback and comments.

Expectation–maximisation model (EM): One of the widely used probabilistic method for estimating answer quality in crowdsourcing problems is the expectation–maximisation model (EM) (Moon, 1996). It regards estimating the reliability of students as a "chicken-and-egg" problem reliant on the other side of the problem, estimating the quality of resources. In the absence of ground truth for both user reliability and resource quality, the first probabilistic model used in this study follows an EM-inspired procedure: (1) set an initial value for all students' reliability, (2) infer resource quality based on current values of assessors' reliabilities and ratings, and (3) update assessors' reliability based on the 'goodness' of their decision compared to the final inferred resource quality.

Trust propagation: Trust propagation models have been successfully employed in various settings such as social networks, commerce, health, and learning (Urena et al., 2019). In the context of peer assessment, this approach is portrayed as a graph with four node types:

students, decision ratings, resources and instructors. The first steps of the trust propagation procedure are similar to those of EM. Each student is first assigned an initial level of trustworthiness. During the decision-making stage, the system estimates the quality of a resource based on the reliability of trustworthy moderators who have evaluated the resource. Then, the assessors' gained reliability score is calculated using the resource's final inferred quality during the user updating stage. However, there is an additional trust propagation stage in which all other users connected to the current assessors receive an updated score from the most reliable and trustworthy ones (Darvishi et al., 2021).

Comment Length: Rather than focusing on reviewer trustworthiness, a large body of natural language processing (NLP) research has sought to assess review quality. Commonly, the length of a comment is used to estimate how much effort assessors put into the assessment (Duret et al., 2018; Xiong & Litman, 2011). In this weighted aggregation model, students who provide a more detailed explanation for their rating are rewarded. For a detailed implementation of this model please refer to (Darvishi et al., 2020).

Relatedness: Although feedback length might reveal student effort, it does not always show how feedback relates to the assessed resource. A comment that refers explicitly to aspects of the resource under evaluation can be more insightful and indicative of critical thinking than a long generic comment. In this text analysis model for peer assessment evaluation, SBERT (Reimers & Gurevych, 2019) is used to generate a semantic vector space for both the comment and the resource, then compute their cosine similarity in that space to measure their relatedness, which is then used as the weight of the assessor's rating to infer the final decision.

Appendix B presents a formal definition for the problem under investigation, followed by eight representative models from four categories of consensus approaches—summary statistic, probabilistic, text analysis and combined models.

Evaluation

We investigated the effectiveness of different consensus models using the following research questions.

RQ2-1: How do commonly used summary statistics infer the quality of student generated content (SGC) in the peer review process?

RQ2-2: How does incorporating probabilistic or text analysis models impact the SGC quality inference?

RQ2-3: Do combinations of the above models improve the performance of the SGC quality inference?

To answer these questions, we collected data from ten courses at The University of Queensland during semester 2 of 2021. Table B1 in Appendix B presents short descriptions of each course and the details for peer assessments regarding the number of students and peer reviews, and Table B2 shows an illustrative dataset. There were a total of 2837 undergraduate students who submitted 60,622 peer assessments on 11,481 resources. Additionally, instructors spot-checked 1017 resources that had received 5856 peer assessments from 1602 students. RiPPLE prioritises a few resources for instructor inspection based on multiple factors such as user feedback, low effectiveness, assessor disagreements, and questionable distractors. Section “Instructor Oversight” delves deeper into the spot-checking algorithm. These spot-checked resources were selected as the test set for evaluation. The peer assessment process in RiPPLE determines whether a student-generated resource is suitable for inclusion in the approved resource repository.

TABLE 1 Comparison between the inferred consensus ratings and the instructors' decisions evaluated with area under the curve (AUC) and accuracy (ACC)

Category	Model	AUC	ACC
Summary statistic	Mean	**0.55**	**0.70**
	Median	0.53	0.69
Probabilistic	Expectation–maximisation	0.58	0.72
	Trust propagation	**0.77**	**0.80**
Text analysis	Length	**0.66**	**0.76**
	Relatedness	0.62	0.74
Combined models	Length × Relatedness	0.68	0.76
	Trust + (Length × Relatedness)	***0.80***	***0.82***

Note: Numbers in bold highlight the best-gained results for each category.
Italic values show the best result for AUC and ACC.

Therefore, we conduct the evaluation at the binary level, where the quality rating = 3 serves as the minimum required inferred rating for approval. This minimum rating is selected based on the rubric (shown in Figure 3a), wherein the Decision section asks:" The overall quality of the resource is: 1 = Poor, 2 = Needs improvement, 3 = Satisfactory, 4 = Great, and 5 = Outstanding". The overall performance of the methods under investigation are shown in Table 1.

Response to RQ2.1: Summary statistics. As baselines, we used the mean and median approaches from summary statistics. The results show that these baselines underperformed and had the lowest AUC among all models, which indicate that the majority of students are easy graders.

Response to RQ2.2: Probabilistic and text analysis methods. The EM method has marginally enhanced the AUC value by re-weighting students' contributions as compared to the baselines. However, the results show that this model was incapable of dealing with data skewness and was still biased towards the majority of overrated assessments. This result is consistent with El Maarry et al. (2015)'s findings, which emphasise that using EM for consensus on data with long-tail distributions may not be appropriate and may facilitate misbehaviour by strategic spammers who provide the most prevalent rating (eg, a high rating here) in their evaluations. In contrast, the trust propagation model substantially improved AUC and ACC when compared to baselines and EM. This finding implies that the graph-based model was more effective than EM at identifying trustworthy student assessors and provided more reliable results. Incorporating linguistic information from comments resulted in better AUC and ACC values as compared to baselines. This improvement implies an association between the additional features from text analysis and the quality of student assessments, and it helps estimate the reliability of reviews and minimises the contribution of strategic spammers.

Response to RQ2.3: Combined model. Although combined models can incorporate many features from preceding models, we only report on the two models that improved performance the most. Interestingly, while we initially assumed that an advanced feature like relatedness (using the BERT model) would be a better predictor of review quality, it did not outperform a simple feature such as the length. We speculate that this could be due to the utilised pre-trained model that often overrates short comments such as "good question." (ie, providing a high relatedness score, eg, >0.7) when compared to MCQ or TRUE/FALSE questions. Nevertheless, the first presented combined model, which integrated length and relatedness, marginally improved AUC compared to the length or relatedness models alone. This model may address the key issue of using the length alone as a proxy for feedback efficacy, resulting in awarded lengthy but ineffective comments. As a result, the Length × Relatedness model could better reflect how much effort and critical thinking assessors put into their evaluation. The second model proved effective as it incorporates features from the previous best-performing models—trust propagation, length and relatedness—resulting in the best AUC and ACC.

In summary, the findings suggest that estimating the trustworthiness of student assessors and the quality of textual feedback offered promising peer assessment outcomes, which could significantly reduce instructors' workload in large-scale assessment tasks. However, a considerable amount of false cases demonstrate the clear need for instructors' oversight during peer assessment.

Feedback on reviews

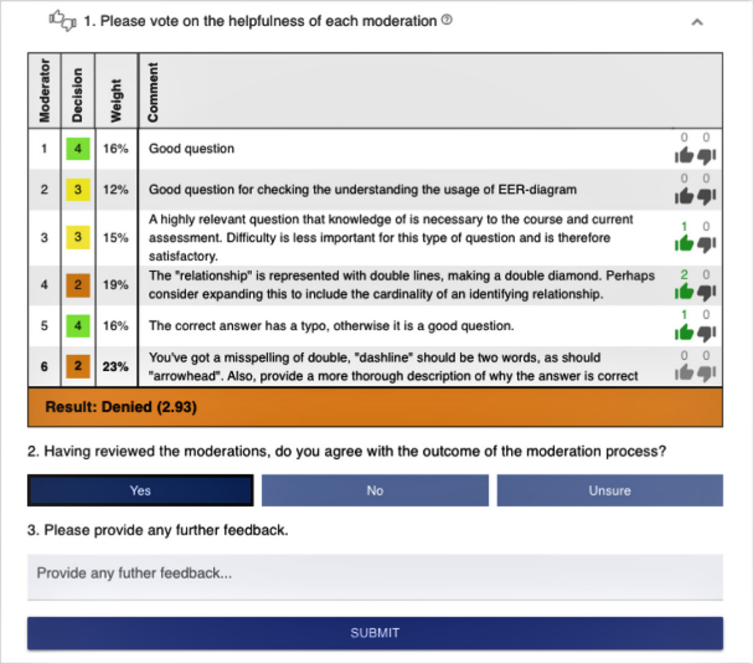
Peer assessment, by design, enables assessors to review, grade and provide feedback on the work of the assessee. This workflow, however, can end up frustrating the assessee or various assessors as they might disagree with the consensus outcome or the provided feedback without being given the opportunity to raise their concerns (Ashenafi, 2017). One possible way to enhance the process and potentially increase its trustworthiness is to close the loop and extend the workflow to enable the assessee and various assessors to review and provide feedback on the work of their peer assessors and the outcome generated by the system. The problem being addressed here also has commonalities with cases where high-stake decisions in domains such as hiring, lending, criminal justice, welfare, and healthcare are automated and made via AI. In both cases, decisions are made by non-experts (students or AI systems), which hinders the trust in the system and may raise concerns for those that are impacted by the decisions (Lockey et al., 2020).

Approach

Prior work on co-regulation (Zheng & Huang, 2016), feedback as a dialogue (Zhu & Carless, 2018) or development of feedback loops (Carless, 2019) provides a strong theoretical framing on how feedback and review loops can be designed in a peer assessment setting. Inspired and informed by these frameworks, crowdsourcing and collaborative learning approaches (Hadwin et al., 2018; Li et al., 2021; Zheng & Huang, 2016), we provide a feedback on reviews strategy for sharing evaluations and inferred decisions with the assessee and various assessors, allowing them to examine and provide feedback on the work of their peer assessors and the system's outcome. In the first step of this approach, the assessee and assessors of the same resource are requested to vote on the helpfulness of each comment and rate the quality of the other assessors' input using like and dislike. This procedure is performed prior to disclosing the results of peer assessment in order to avoid user bias based on the outcome. Following the sharing of the result, the next step is for the participants to declare whether or not they agree with the outcome. This enables them to reconsider their previous decision and confirm or change their initial opinion. Finally, they are given a chance to anonymously offer additional comments to share their experience and possible concerns about the peer assessment process. Figure 5 depicts RiPPLE's feedback on reviews interface.

Evaluation

The following research questions guide our exploration of students' engagement with the feedback on reviews interface and the effectiveness of our approach.



1. Please vote on the helpfulness of each moderation

Moderator	Decision	Weight	Comment	
1	4	16%	Good question	0 👍 0 👎
2	3	12%	Good question for checking the understanding the usage of EER-diagram	0 👍 0 👎
3	3	15%	A highly relevant question that knowledge of is necessary to the course and current assessment. Difficulty is less important for this type of question and is therefore satisfactory.	1 👍 0 👎
4	2	19%	The "relationship" is represented with double lines, making a double diamond. Perhaps consider expanding this to include the cardinality of an identifying relationship.	2 👍 0 👎
5	4	16%	The correct answer has a typo, otherwise it is a good question.	1 👍 0 👎
6	2	23%	You've got a misspelling of double, "dashline" should be two words, as should "arrowhead". Also, provide a more thorough description of why the answer is correct	0 👍 0 👎

Result: Denied (2.93)

2. Having reviewed the moderations, do you agree with the outcome of the moderation process?

[✓ Yes] No Unsure

3. Please provide any further feedback.

Provide any futher feedback...

SUBMIT

FIGURE 5 Example of feedback on reviews interface in RiPPLE, which shows six students' decisions yielded to rejection of the resource under moderation, three comments received like after the result and feedback of the peer assessment were shared with students, and the current user in the feedback on reviews interface declares that they agree with the outcome

> **RQ3-1:** To what extent do students trust the system and agree with the outcomes given by the system?
>
> **RQ3-2:** How likely is it for an assessor to change their opinion based on the reviews of others?
>
> **RQ3-3:** What are the common topics discussed by the authors and assessors in their additional comments?

To answer these questions, we collected data from 1,348 users who engaged in 12,747 instances of feedback on reviews based on the interface shown in Figure 5, which captures students' feedback after the results of the peer assessment were shared with them. Table C1 in Appendix C shows an illustrative dataset. Providing feedback on reviews is not mandated in the peer assessment system, and participation is voluntary. Our data showed a significant difference between the number of peer reviews submitted per student (medians = 16, Q1 = 12 and Q3 = 30) and their number of feedback on reviews (medians = 4, Q1 = 1 and Q3 = 11), where Q1-Lower Quartile is 25th, and Q3-Upper Quartile is 75th percentile. Table 2 provides an overview of the feedback on review contributions by drilling down into the users' likes/dislikes, votes and additional comments.

Response to RQ3-1: Trust. As shown in Table 2, only around 2% of the responses disagreed with the outcomes of the peer assessment process and less than 4% were unsure. In contrast, more than 80% of the responses agreed with the outcomes of the peer assessment process. These stats suggest that the general trust of the students in the system is quite high. We note that around 13% of the users who submitted a feedback on reviews contribution did not vote and therefore, we do not know whether or not they had trust in the system.

TABLE 2 Users' vote per peer assessment outcome

User	Assessment outcome	User decision	#Likes	#Dislikes	User vote					#Comments
					Not voted	Disagree	Unsure	Agree	Total	
Authors	Reject	NA	237	52	11 (0.4%)	23 (0.7%)	12 (0.4%)	64 (2%)	110 (3.5%)	54
	Approve	NA	8448	254	437 (13.9%)	43 (1.4%)	86 (2.7%)	2466 (78.5%)	3032 (96.5%)	815
	Total		8685	306	448 (14.3%)	66 (2.1%)	98 (3.1%)	2530 (80.5%)	3142	869
Assessors	Reject	Reject	417	147	37 (0.4%)	8 (0.1%)	7 (0.1%)	229 (2.4%)	281 (2.9%)	42
		Approve	390	41	27 (0.3%)	18 (0.2%)	22 (0.2%)	146 (1.5%)	213 (2.2%)	35
	Approve	Reject	419	233	40 (0.4%)	110 (1.1%)	54 (0.6%)	133 (1.4%)	337 (3.5%)	96
		Approve	17,086	702	1199 (12.5%)	75 (0.8%)	269 (2.8%)	7231 (75.3%)	8774 (91.4%)	701
	Total		18,312	1123	1303 (13.6%)	211 (2.2%)	352 (3.7%)	7739 (80.6%)	9605	874

Response to RQ3-2: Change opinion. In four cases, assessors showed a change in their initial decision after reviewing feedback: (1) Outcome and their decision were both to reject a resource, but they disagreed with the outcome (0.1%). This scenario was rare, with only 8 out of 281 cases with similar conditions (probability = 2.8%). (2) Outcome was to reject a resource, but the decision was to approve; however, assessors agreed with the outcome (1.5%). This scenario was relatively common with 146 out of 213 similar condition cases (probability = 68.5%). (3) Outcome was to approve a resource, but the decision was to reject; however, assessors eventually agreed with the outcome (1.4%). Compared to the previous scenario, assessors were less likely to change their minds and agree with the outcome in this case, with 133 out of 337 similar condition cases (probability = 39.4%). Finally, (4) outcome and their decision were both to approve a resource, but assessors disagreed with the outcome (0.8%). This scenario is the least common possibility with 75 out of 8744 similar condition cases (probability = 0.9%). All in all, assessors changed their opinions in 3.8% of cases when they participated in the feedback on reviews process, whereas it would be 1.3% if all submitted assessments were considered.

Response to RQ3-3: Common topics. To answer this question, we randomly sampled and manually coded 10% (175) of the user comments to detect recurring topics. One of the authors and a research assistant independently coded the comments. We used Roberts et al. (2019)'s guidelines on codebook generation. The preliminary codes were deduced and applied to the comments based on the research question. This initial codebook was then iterated several times through an inductive process, where additional codes were formed based on the data before the coders agreed on the final version. To achieve appropriate levels of inter-rater agreement, we concluded the coding using a test–retest technique. The inter-rater agreement across all codes was 93.5% with a Cohen's kappa coefficient of 0.84, indicating a strong agreement. The coders discussed cases of disagreements between them and resolved conflicts. Our analysis revealed that users primarily provided feedback on five topics: their peers (either assessees or other assessors), themselves, the outcome, the system, and the resource under assessment. Table C2 in Appendix C reports rates of occurrence as well as positive and negative example comments for assessees and assessors on each of these five topics. We note that a comment could have been tagged by multiple topics. Therefore, the sum of the rates of occurrences exceeds 100% for both assessees and assessors. Results suggest that the assessed resource's content is the central point of discussion for both assessees (90.8%) and assessors (76.1%). However, whereas assessees concentrated on themselves (37.9%) as the next most discussed topic, assessors focused on their peers (39.8%). The majority of assessees' positive comments reflect their development in feedback literacy by taking up on feedback and appreciating their peers' points of view and suggestions (Carless & Boud, 2018). On the other hand, while some negative comments addressed the unfairness of outcome, most of them complained about generic, non-constructive and unhelpful textual feedback. The comments on the system topic enable RiPPLE's designer and development team to resolve reported issues and improve the useability of the system.

In summary, while 80% of students trusted the peer assessment process and agreed with the outcome, feedback on reviews allows peer assessors and assessees to review feedback and engage in dialogue with one another, change their initial opinion based on others' decisions, and express disagreement with the outcome. It also allows them to express additional concerns about their peers' contributions, the content of the evaluated resource, and the system's functionality, which can assist the system and instructors in identifying issues raised by the students.

Instructor oversight

An important element that gives credibility to peer assessment is that instructors, as the experts, are overseeing the process. However, despite the plethora of previous work on peer assessment, methods and processes on how instructors can best oversee the peer assessment process are underexplored. One simple approach would be for instructors to review all peer-reviewed cases; however, this approach requires a significant instructor effort and would be unrealistic to implement in large classes. An alternative approach that has received some attention is the development of spot-checking mechanisms (Lee et al., 2018; Wang et al., 2018) that aim to identify a small subset of peer assessment outcomes that would benefit the most by being checked by instructors (Yang et al., 2019). This is a promising approach that can improve the trustworthiness of peer assessment systems while optimally using the expertise of time-poor instructors. However, much of the existing work on spot-checking has had a theoretical nature and has only been evaluated on simulated or offline data sets rather than empirical evaluation in peer assessment systems (Han et al., 2020; Wang et al., 2020).

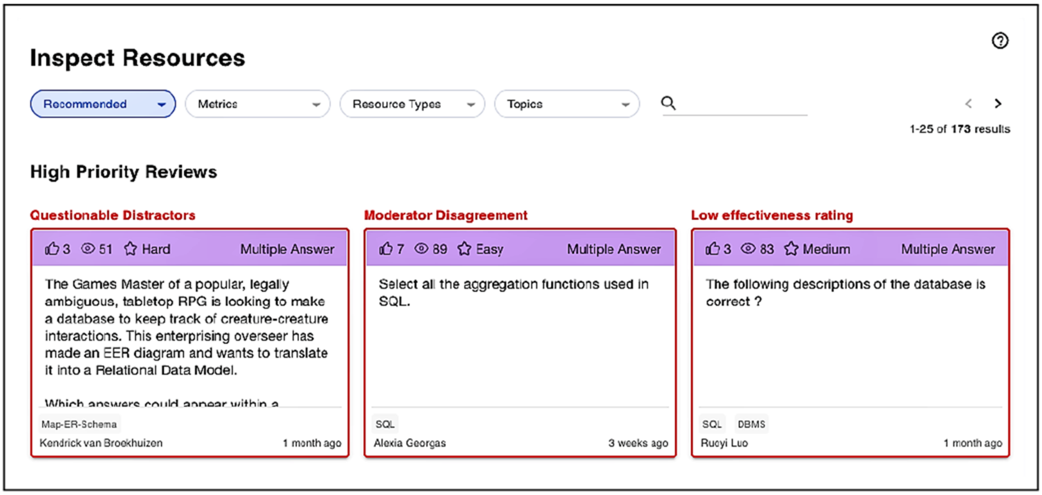
Approach

Informed by spot-checking (Wang et al., 2020) and active learning (Lee et al., 2018) approaches, we developed a set of learning analytics that help instructors to effectively use their time to review the most controversial cases. RiPPLE incorporates a spot-checking algorithm that employs a blend of human-driven and data-driven metrics for identifying resources that would benefit the most from being reviewed by instructors. Human-driven metrics include outcome disagreement (assessors/authors disagree with the final decision on a resource), low efficacy (a high proportion of downvotes (dislikes) relative to upvotes (likes)), and reports (resources reported as unsuitable or having incorrect answers by students). Data-driven metrics take into account assessment items with questionable distractors, in which the popular answer is not the one provided by the author. When flagging a resource for spot checking, RiPPLE uses absolute and relative points of comparison to help instructors make sense of why a resources has been flagged for review. Table D2 in Appendix D shows examples of metrics used for spot-checking along with their descriptions, explanations and examples. Each resource receives a risk score between 0 to 5 on each of the metrics where 5 presents the highest level of risk or need for expert review.

We computed an overall ranking score for resources based on their priority to require instructor judgement. To obtain a ranking score, we made use of stochastic dominance (Levy & Robinson, 2006) from decision theory where one set of outcomes (set of metric risk scores for a resource) could be considered superior to another set of outcomes (set of metric risk scores for another resource). The aim was to order the resources in such a way that one resource has a lower priority than another if their distribution of outcomes is both smaller on average and less variable, while resources with a higher priority have larger scores on average and more variability. For obtaining our function, we considered the following two criteria, which are related to the first-order and second-order stochastic dominance: (1) The overall ranking score for a resource r dominates the overall ranking score for resource r' if each metric risk scores of resource r dominates that of scores of resource r'. (2) If the average of the metric risk scores of resource r is equal to that of resource r', then the resource with the higher standard deviation across metric scores will have a higher overall score. This criterion would give a higher overall ranking score to resources that have a high score on some metrics over those that have a medium score across all metrics. Figure 6 shows the RiPPLE interface in which the system prioritises items based the metrics described in Table D2.

Evaluation

Our evaluation of the spot checking approach was guided by the following research questions.



Inspect Resources

1-25 of 173 results

High Priority Reviews

Questionable Distractors

👍 3 👁 51 ☆ Hard Multiple Answer

The Games Master of a popular, legally ambiguous, tabletop RPG is looking to make a database to keep track of creature-creature interactions. This enterprising overseer has made an EER diagram and wants to translate it into a Relational Data Model.

Which answers could appear within a

Map-ER-Schema

Kendrick van Broekhuizen 1 month ago

Moderator Disagreement

👍 7 👁 89 ☆ Easy Multiple Answer

Select all the aggregation functions used in SQL.

SQL

Alexia Georgas 3 weeks ago

Low effectiveness rating

👍 3 👁 83 ☆ Medium Multiple Answer

The following descriptions of the database is correct ?

SQL DBMS

Rueyi Luo 1 month ago

FIGURE 6 Example of instructor oversight in RiPPLE

TABLE 3 Summary of instructors' oversight

(a) Instructors' actions and decisions

	Instructor action		
Final decision	**Flag cleared**	**Revised**	**Total**
Reject	17	521	538
Approve	410	380	790
Total #resources	427	901	1328

(b) Instructors' actions based on number of flags

	Number of flag per item			
Instructor action	**1**	**2**	**3**	**4**
Flag cleared	405	22	0	0
Revised	781	106	10	4
Total #flags	1186	256	30	16

> **RQ4-1:** How have instructors engaged with the spot-checking algorithm?

> **RQ4-2:** How effective was each metric in urging instructors to revise the outcome of a resource?

To answer these questions, we collected and analysed data on 1328 cases where instructors reviewed an item that the system prioritised for inspection using one of the metrics mentioned earlier. Table D1 in Appendix D shows an illustrative dataset. Our data from ten different courses revealed that instructors had been generally involved and spot-checked less than 2% of all student submissions, with the median of 1.97%, lower quartile Q1 = 1.46%, and upper quartile Q3 = 7.28%.

TABLE 4 Instructors' action based on the type of metric

Instructor action	Low effectiveness	Users' reports	Assessors' disagreement	Questionable distractors
Flag cleared	10.7%	18.1%	22.3%	44.6%
Revised	89.3%	81.9%	77.7%	55.4%
Total #flags	56	298	524	610

Response to RQ4-1. Table 3a shows that instructors cleared flag on 32% (ie, 427 out of 1328) of the prioritised resources by the assisted spot-checking, which means that these items did not need instructors' intervention. However, this table also shows that instructors revised 901 out of 1328 of items prioritised by the system, indicating the spot-checking algorithm's 68% efficacy in detecting resources that benefit from instructors oversight. Also, instructors rejected 58% (ie, 521 out of 901) of the revised items, while the remaining 42% of approved cases were also revised by providing a new grade. A number of resources were flagged by more than one metrics. Table 3b shows that the chance of revising a resource has increased by the number of flags from 66% in 1-flagged items to 83% in 2-flagged items and 100% in 3- or 4-flagged items.

Response to RQ4-2. According to Table 4, questionable distractors and assessors' disagreement appear to have been front-runner metrics used by the system to flag potential undependable resources with 610 and 524 cases, respectively. They are followed by users' reports and low effectiveness with 298 and 56 flagged items, respectively.

Regarding the number of ratings provided by instructors, assessors' disagreement had the most cases (ie, 407), but 117 flagged cases by this metric were cleared which makes its efficacy 77.7%. However, low effectiveness, despite its low contribution in detecting problematic resources, at 89.3% (ie, 50 out of 56 cases) was the most successful metric in encouraging instructors to revise and submit a new rating. Users' reports had the second-highest efficacy rate of 81.9% with 244 re-rated items out of 298 cases. While questionable distractor had the highest number of flagged items, 44.6% of items flagged by this metric were eventually cleared by instructors, making it the least effective one with a 55.4% efficacy. This suggests that students simply misunderstood a question in many questionable distractor cases due to the question difficulty level.

In summary, our proposed analytics and recommender system for identifying resources that need instructor attention seems to be effective as a large portion of the recommended resources receive a revision grade from instructors.

DISCUSSION AND CONCLUSION

In this work, we presented AI-assisted and analytic approaches to address some of the main concerns associated with the use of peer assessment systems. By doing so, we aimed to give instructors, students and institutes a firmer belief in the competence and reliability of peer assessment systems, which based on the definition by (Robinson, 1996, p. 3), can contribute to their trustworthiness. In particular, our results provide evidence that our proposed approaches (1) enable students to write lengthier and more helpful feedback comments, (2) infer student grading reliability such that the results are more accurate than current baseline model, (3) enable students to provide peer feedback over peer reviews and raise concerns when needed and (4) enable instructors to incorporate system's recommendations to more optimally identify cases that need instructor oversight. The findings point to a future in which peer assessment may be incorporated more frequently and reliably to facilitate learning at scale.

Implications

The paper has two important implications for learning analytics and AI in education. First, this paper gives researchers and practitioners a novel systematic approach to incorporating advances in AI-driven learning analytics by having a strong grounding in a theoretical model of relevant education or learning processes. Specifically, the paper—through the conceptualisation of the four phases of peer-review process—demonstrated how a theoretical model can be used to structure the program of research, development, deployment, and evaluation by addressing a problem (ie, trust in a peer-review system) that may emerge in practice. This is an important contribution that complements existing calls for integration of theory and design with computational methods of AI and data science in learning analytics (Gašević et al., 2017; Martinez-Maldonado et al., 2021). Second, the studies reported in the paper provide fresh empirical insights that can inform the development of future AI-driven learning analytic systems that seek to enhance trustworthiness of peer-review. To this end, the paper advances existing research that mostly focuses on design proposals (Er et al., 2021) for the use of learning analytics to support peer feedback and make use of learning analytics to evaluate peer assessment approaches (Kulkarni et al., 2015).

Limitations

We see two main limitations to the current study. First, while the applied methods are presumed to be context and domain agnostic, this study only examined data from a single system with a particular context of peer assessment, which might lead to potential bias in findings. Our future work aims to replicate the presented studies in other peer assessment systems. Second, we have approached increasing trustworthiness by aiming to address some of the main concerns associated with using peer assessment systems. However, in this work, we have not directly measured the impact of our approaches on users' perceptions of trustworthiness. Future work aims to incorporate self-reports and to conduct focus groups for triangulation of findings and to evaluate the effectiveness of our approach on increasing trust.

FATE concerns

While we have shown examples of how AI and analytics can assist in addressing concerns with peer assessment, there are also increasing concerns about the FATE (Fairness, Accountability, Transparency, and Ethics) of AI-based systems (Shin, 2020; Darvishi et al., 2021). Here we provide one example of a potential pitfall for each of the four peer assessment processes we have studied to indicate that further exploration and evaluation are required before we aggressively incorporate AI and analytics in peer assessment.

Individual reviews

The use of NLP algorithms seems effective in helping students develop algorithmic literacy (Koenig, 2020; Long & Magerko, 2020; Darvishi et al., 2022); however, inaccurate prompts and recommendations by the system (eg, asking a reviewer to add an explicit suggestion, while the review already has one) may lead to students and instructors losing their trust in the system. An interesting future direction is to explore best practices for developing trustworthy recommender systems (Hassan, 2019).

Assigning grades

We have demonstrated that the use of probabilistic and text-based models can improve the accuracy of peer grading in comparison to baseline summary statistics. However, grades inferred via these models are also less explainable and interpretable by students and instructors, which itself can increase distrust in the system. An interesting direction for future work, which is aligned with the broader work on the emerging field of explainable AI (Arrieta et al., 2020), is to develop methods that are both explainable but are also highly accurate.

Feedback on reviews

Providing the ability for students to provide *feedback on reviews* of each other by rating the assessors and raising concerns seems like a promising approach to increase trust in peer assessment systems (Ahmad & Bull, 2008; Bodily et al., 2018). However, these additions introduce the expectation in students that instructors will closely examine and act on their feedback, when appropriate, to change results. This itself can be a source of frustration for both instructors (in terms of additional work) and students (in terms of unmet expectations), leading to less trust and confidence in the system. An interesting direction for future work is to explore feedback on reviews processes that empower students in peer assessment systems without directly increasing instructors' workload.

Instructor oversight

Spot-checking algorithms can assist instructors to review only a small portion of the assessments. However, there is a concern that the selection process may be prone to data and algorithmic bias (Baker & Hawn, 2021), which may generate underlying discrimination that might unfairly benefit or harm certain groups of students. An interesting direction for future work, which is aligned with the broader work on human-centred AI, is to evaluate the fairness of our approach against various demographic groups using methods that consider fairness measures (eg, [Gardner et al., 2019]) and to develop new approaches that are less prone to user or algorithmic bias.

ACKNOWLEDGEMENTS

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. Open access publishing facilitated by The University of Queensland, as part of the Wiley - The University of Queensland agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this paper.

ETHICS STATEMENT

Approval from our Human Research Ethics Committee (\#2018000125) was received for conducting these studies.

DATA AVAILABILITY STATEMENT

Publishing and including students' data is restricted by the ethical approval from our Human Research Ethics Committee. Instead, we have published and shared the computational implementations and developed codes of the proposed methods in a GitHub repository with illustrative datasets in appendices, which can help the implementation for interested researchers

ORCID

Ali Darvishi https://orcid.org/0000-0002-7025-9259

ENDNOTES

[1] The source codes (in R, Python, and Jupyter notebook) are available at: https://github.com/ali-darvishi/BJET_Trustworthy-Peer-Assessment.

[2] Approval from our Human Research Ethics Committee (#2018000125) was received for conducting these studies.

REFERENCES

Abdi, S., Khosravi, H., & Sadiq, S. (2020). Modelling learners in crowdsourcing educational systems. In *International Conference on Artificial Intelligence in Education* (pp. 3–9). Springer.

Abdi, S., Khosravi, H., Sadiq, S., & Darvishi, A. (2021). Open learner models for multi-activity educational systems. *Artificial Intelligence in Education*, 11–17. https://doi.org/10.1007/978-3-030-78270-2_2

Ahmad, N., & Bull, S. (2008). Do students trust their open learner models? In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 255–258). Springer.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Ashenafi, M. M. (2017). Peer-assessment in higher education-twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education*, *42*, 226–251.

Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between student's effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review*, *32*, 1–24.

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *31*, 1–41.

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: a systematic review. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 41–50). Association for Computing Machinery.

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

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment & Evaluation in Higher Education*, *44*, 705–714.

Carless, D. (2020). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*, 1469787420945845. https://doi.org/10.1177/1469787420945845

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*, 1315–1325.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*, 73–84.

Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learner sourcing to inform design loop adaptivity. In *European Conference on Technology Enhanced Learning* (pp. 332–346). Springer.

Darvishi, A., Khosravi, H., & Sadiq, S. (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* (pp. 139–150). Association for Computing Machinery.

Darvishi, A., Khosravi, H., Sadiq, S., & Weber, B. (2021). Neurophysiological Measurements in Higher Education: A Systematic Literature Review. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00256-0

Darvishi, A., Khosravi, H., Abdi, S., Sadiq, S., & Gašević, D. (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. https://doi.org/10.1145/3491140.3528265

De Alfaro, L., & Shavlovsky, M. (2014). CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education* (pp. 415–420). Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duret, D., Christley, R., Denny, P., & Senior, A. (2018). Collaborative learning with peerwise. *Research in Learning Technology*, *26*, 1–13.

El Maarry, K., Güntzer, U., & Balke, W.-T. (2015). A majority of wrongs doesn't make it right-On crowdsourcing quality for skewed domain tasks. In *International Conference on Web Information Systems Engineering* (pp. 293–308). Springer, Cham.

Er, E., Dimitriadis, Y., & Gašević, D. (2021). Collaborative peer feedback and learning analytics: Theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, *46*, 169–190.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225–234). Association for Computing Machinery.

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice*, *3*, 63–78.

Gyamfi, G., Hanna, B. E., & Khosravi, H. (2021). The effects of rubrics on evaluative judgement: A randomised controlled experiment. *Assessment & Evaluation in Higher Education*, *47*(1), 126–143. https://doi.org/10.1080/02602938.2021.1887081

Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In *Handbook of self-regulation of learning and performance* (pp. 83–106). Routledge.

Han, Y., Wu, W., Yan, Y., & Zhang, L. (2020). Human-machine hybrid peer grading in SPOCs. *IEEE Access*, *8*, 220922–220934.

Hassan, T. (2019). Trust and trustworthiness in social recommender systems. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 529–532). Association for Computing Machinery.

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development*, *38*, 1401–1416.

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology*, *49*, 101–110.

Jansen, R. S., Van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (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*, 100292.

Joyner, D. A. (2017). Scaling expert feedback: Two case studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 71–80). Association for Computing Machinery.

Kao, G. Y.-M. (2013). Enhancing the quality of peer review by reducing student "free riding": Peer assessment with positive interdependence. *British Journal of Educational Technology*, *44*, 112–124.

Khosravi, H., Kitto, K., & Joseph, W. (2019). Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, *6*, 91–105.

Khosravi, H., Demartini, G., Sadiq, S., & Gasevic, D. (2021). Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21)* (pp. 32–42). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448143

Khosravi, H., Gyamfi, G., Hanna, B. E., Lodge, J., & Abdi, S. (2021). Bridging the gap between theory and empirical research in evaluative judgment. *Journal of Learning Analytics*, *8*(3), 117–132. https://doi.org/10.18608/jla.2021.7206

Koenig, A. (2020). The algorithms know me and i know them: Using student journals to uncover algorithmic literacy awareness. *Computers and Composition*, *58*, 102611.

Kulkarni, C. E., Bernstein, M. S., & Klemmer, S. R. (2015). PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 75–84). Association for Computing Machinery.

Lahza, H., Khosravi, H., Demartini, G., & Gasevic, D. (2022). Effects of technological interventions for self-regulation: A control experiment in learnersourcing. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery. https://doi.org/10.1145/3506860.3506911

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. *Frontiers in Psychology*, *4*, 863.

Lee, W., Huang, C. H., Chang, C. W., Wu, M. K. D., Chuang, K. T., Yang, P. A. and Hsieh, C. C. (2018) Effective quality assurance for data labels through crowdsourcing and domain expert collaboration. In *21st International Conference on Extending Database Technology, EDBT 2018* (pp. 646–649). OpenProceedings.org.

Levy, H., & Robinson, M. (2006). *Stochastic dominance: Investment decision making under uncertainty* (Vol. *34*). Springer.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*, 193–211.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*, 525–536.

Li, Y., Jin, X., Hu, Q., Jiang, Q., Zhao, W., & Oubibi, M. (2021). An empirical study on the influence of co-regulation on deep learning under crowdsourcing knowledge construction. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 526–530). IEEE.

Lin, J.-W. (2018). Effects of an online team project-based learning environment with group awareness and peer evaluation on socially shared regulation of learning and self-regulated learning. *Behaviour & Information Technology*, *37*, 445–461.

Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*, 279–290.

Lockey, S., Gillespie, N. and Curtis, C. (2020) *Trust in artificial intelligence: Australian insights*. The University of Queensland and KPMG. https://doi.org/10.14264/b32f129

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery.

Lundstrom, K., & Baker, W. (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*, 30–43.

Manso-Vázquez, M., & Llamas-Nistal, M. (2015). A monitoring system to ease self-regulated learning processes. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *10*, 52–59.

Martinez-Maldonado, R., Gaševic, D., Echeverria, V., Fernandez Nieto, G., Swiecki, Z., & Buckingham Shum, S. (2021). What do you mean by collaboration analytics? A conceptual model. *Journal of Learning Analytics*, *8*, 126–153.

Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226–245.

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

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (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)* (pp. 588–593). Association for Computational Linguistics (ACL).

Negi, S., Asooja, K., Mehrotra, S., & Buitelaar, P. (2016). A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 170–178). Association for Computational Linguistics.

Nejad, A. M., & Mahfoodh, O. H. A. (2019). Assessment of oral presentations: Effectiveness of self-, peer-, and teacher assessments. *International Journal of Instruction*, *12*, 615–632.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*, 375–401.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*, 102–122.

Paré, D. E., & Joordens, S. (2008). Peering into large lectures: Examining peer and expert mark agreement using peerscholar, an online peer assessment tool. *Journal of Computer Assisted Learning*, *24*, 526–540.

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, *43*, 591–614.

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

Polisda, Y. (2017). Peer review: A strategy to improve students' academic essay writings. *English Franca: Academic Journal of English Language and Education*, *1*, 45–60.

Purchase, H., & Hamer, J. (2018). Peer-review in practice: Eight years of Aropä. *Assessment & Evaluation in Higher Education*, *43*, 1146–1165.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, *27*, 534–581.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics.

Roberts, K., Dowell, A., & Nie, J.-B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Medical Research Methodology*, *19*, 1–8.

Robinson, S. L. (1996). Trust and breach of the psychological contract. *Administrative Science Quarterly*, *41*, 574–599.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*, 541–565.

Shnayder, V., Agarwal, A., Frongillo, R., & Parkes, D. C. (2016). Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 179–196). Association for Computing Machinery.

Shoham, N., & Pitman, A. (2021). Open versus blind peer review: Is anonymity better than transparency? *BJPsych Advances*, *27*, 247–254.

Sommers, N. (1982). Responding to student writing. *College Composition and Communication*, *33*, 148–156.

Sridharan, B., Tai, J., & Boud, D. (2019). Does the use of summative peer assessment in collaborative group work inhibit good judgement? *Higher Education*, *77*, 853–870.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*, 467–481.

Tao, D., Cheng, J., Yu, Z., Yue, K., & Wang, L. (2018). Domain-weighted majority voting for crowdsourcing. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 163–174.

Topping, K. (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* (pp. 55–87). Springer, Dordrecht.

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*, 20–27.

Topping, K. J. (2010). Peers as a source of formative assessment. In *Handbook of formative assessment* (pp. 73–86). Routledge.

Urena, R., Kou, G., Dong, Y., Chiclana, F., & Herrera-Viedma, E. (2019). A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. *Information Sciences*, *478*, 461–475.

Wang, W., An, B. and Jiang, Y. (2018) Optimal spot-checking for improving evaluation accuracy of peer grading systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. *32*, No. 1). AAAI Press.

Wang, W., An, B., & Jiang, Y. (2020). Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems. *IEEE Transactions on Computational Social Systems*, *7*, 940–955.

Wind, D. K., Jørgensen, R. M., & Hansen, S. L. (2018). Peer feedback with peergrade. In *ICEL 2018 13th International Conference on e-Learning* (p. 184). Academic Conferences and Publishing Limited.

Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (pp. 96–101). Association for Computing Machinery.

Xiong, W., & Litman, D. (2011). Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 502–507). Association for Computational Linguistics.

Yang, M., Tai, M., & Lim, C. P. (2016). The role of e-portfolios in supporting productive learning. *British Journal of Educational Technology*, *47*, 1276–1286.

Yang, T.-Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active learning for student affect detection. In *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019* (pp. 208–217). Université du Québec; Polytechnique Montréal.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, *143*, 804–824.

Yu, F.-Y., & Wu, C.-P. (2011). Different identity revelation modes in an online peer-assessment learning environment: Effects on perceptions toward assessors, classroom climate and learning activities. *Computers & Education*, *57*, 2167–2177.

Zheng, L., & Huang, R. (2016). The effects of sentiments and co-regulation on group performance in computer supported collaborative learning. *The Internet and Higher Education*, *28*, 59–67.

Zhu, Q., & Carless, D. (2018). Dialogue within peer feedback processes: Clarification and negotiation of meaning. *Higher Education Research & Development*, *37*, 883–897.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy.* American Psychological Association.

Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior*, *124*, 106924.

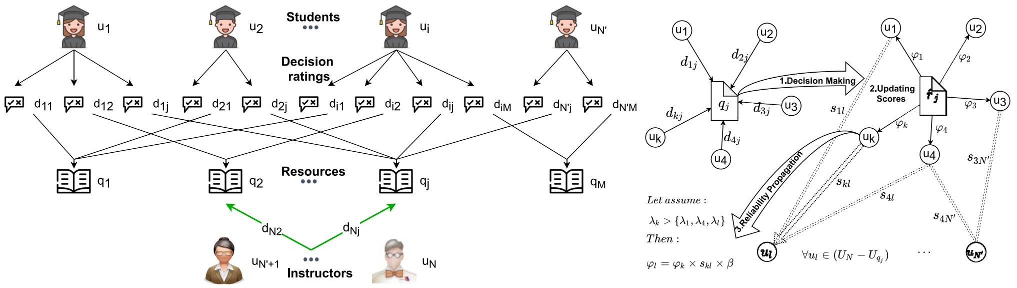
How to cite this article: Darvishi, A., Khosravi, H., Sadiq, S. & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, 53, 844–875. https://doi.org/10.1111/bjet.13233

APPENDIX A

SUMMARY OF DATASET USED IN *INDIVIDUAL REVIEWS*

TABLE A1 Overview of the experimental groups in *individual reviews*

Course	Group	#Students	#Resources	#Peer reviews
NEUR	Control	117	703	1290
	Experiment	117	573	1247
INFS	Control	70	165	342
	Experiment	70	216	304
Total		374	1657	3183



(a) Moderation graph

(b) Sample propagation network

FIGURE A1 Graph-based trust propagation. (a) Moderation graph with four kinds of node—Students, decision ratings, resources, and instructors, and (b) main steps in the propagation network model

APPENDIX B

INFERENCE MODELS AND DATASET USED IN *ASSIGNING GRADES*

This Appendix presents a formal definition and notation for the problem under investigation, followed by eight representative models from four categories of consensus approaches—summary statistic, probabilistic, text analysis, and combined models.

Problem definition

Given a non-moderated resource q_j and a set of students $\{u_1 \dots u_k\}$ who provided decision ratings $d_{1j}, \dots, d_{kj}$ accompanied with comments $c_{1j}, \dots, c_{kj}$, infer the quality of q_j denoted as $\hat{r}_j$..

Summary statistics

Mean. This is a simple consensus approach based on summary statistics: $\hat{r}_j = \frac{\sum_{i=1}^{k} d_{ij}}{k}$.

Median. The median is usually situated between the mean and the mode in skewed normal distributions: $\hat{r}_j = \text{Median}(u_1, \dots u_k)$.

Probabilistic

Expectation-Maximisation. In the absence of ground truth, resource q_j quality $(\hat{r}_j)$ is inferred based on current values of student reliability scores $\lambda_1, \dots \lambda_k$ and, ratings $d_1, \dots d_k$ on q_j; and then reliability of the students $\lambda_1, \dots \lambda_k$ are updated using Equation 1 as follows:

$$\hat{r}_j = \frac{\sum_{i=1}^{k} \lambda_i \times d_{ij}}{\sum_{i=1}^{k} \lambda_i}, \quad \lambda_i = \lambda_i + f_{ij}^{R}\left(d_{ij}, \hat{r}_j\right) \tag{1}$$

where $f_{ij}^{R}\left(d_{ij}, \hat{r}_j\right) = \frac{2\delta e^{-\left(d_{ij} - \hat{r}_j\right)^2 / \left(2\sigma^2\right)} - \delta}{2\sigma\sqrt{2\pi}}$ determines the 'goodness' of d_{ij} based on $\hat{r}_j$ as the height of a Gaussian function at value $\left(d_{ij} - \hat{r}_j\right)$ with centre 0, standard deviation $\sigma = .7$ and peak $\delta = 100$.

Graph-based trust propagation. Figure A1a shows the moderation graph consisting of four nodes—students, decision ratings, resources, and instructors. The trust propagation model has three main stages: decision-making, updating scores, and reliability propagation. Similar to EM, in the decision-making and updating score stages, students' ratings

TABLE B1 Short descriptions of each course and the number of students and peer reviews

Course code	Description	School	#Students	#Peer reviews
MEDI	Ethics and Professional Practice	Medical School	415	6467
INFS	Introduction to Information Systems	Info Tech & Elec Engineering	379	4951
NEUR	The Brain and Behavioural Sciences	Psychology School	554	20,939
AGRC	Applied Mathematics & Statistics	Agriculture Food Sciences	292	11,462
CRIM	Introduction to Criminal Justice	Social Science School	160	2492
ECON	The Macroeconomy	Economics School	256	4510
PHRM	Quality Use of Medicines	Pharmacy School	227	2541
COMP	Artificial Intelligence	Info Tech & Elec Engineering	301	3543
NUTR	Nutrition & Exercise	Human Movement & Nutrition Sci	195	3051
ANIM	Wildlife Technologies	Agriculture Food Sciences	58	666

TABLE B2 An illustrative dataset used for *assigning grade*, consisting of 4 resources assessed by 10 students. Resources q1, q2, and q3 have been mostly received high ratings (ie, decision >3), but resource q4 has controversial decisions ranging from 2 (reject), 3 (satisfactory) to 5 (outstanding). Comments length in words and relatedness score (between [0, 1]) were also used to measure textual feedback quality

Resource ID	User ID	Decision	Final comment	Comment length	Relatedness score
q1	u1	4	Very effective resource for learning the basics of the PACE framework	11	0.440993
q1	u2	4	I like the explanation, this is a good question for my own review of the content	16	0.200965
q1	u4	5	Very helpful, great revision tool for remembering the PACE framework	10	0.404459
q1	u10	4	This question is a good review of the PACE framework and a good way to test your knowledge of the framework. The answer explanation is appropriate and helpful	28	0.537852
q2	u1	4	Great question to develop critical thinking skills. It was difficult trying to workout which of the two last answers were correct	21	0.344651
q2	u2	4	This is a great question that tested my recall on the PACE model, in a way which required me to think about particularly between C and D	27	0.430061
q2	u10	4	This question is a good review of the PACE framework. I like how each answer choice is a different part of PACE and they are each covered in the question explanation	31	0.380155
q2	u8	4	I really like that the author used an example to test our understanding of the PACE framework of escalating	19	0.577151
q3	u1	4	Solid question, I think it has strong potential as an exam question	12	0.19901
q3	u3	5	I like how you have discussed the topic of risks and threats in healthcare, great work	16	0.501467
q3	u4	5	Very helpful, very good tool for revision of direct content from lecture slides	13	0.043853
q3	u5	3	Good resource to revise types of risks	7	0.499096
q4	u6	5	Concise and accessible notes for understanding the components of shared decision making and examples on how to put these into practise	21	0.459132
q4	u7	3	good resource added at the bottom allows students to further read if they choose to, adding explanations helped give context	20	0.257424
q4	u8	2	Very superficial resource. I'm not certain what the take home is in terms of KLIs. I'm not sure the formate chosen is the best way to transmitting this information	29	0.36531
q4	u9	3	If you could put which week this information was relevant to, it would help those going back to revise the material:)	22	0.223229
q4	u5	3	Good layout and helpful for revising. May have been more helpful as a practice question	15	0.290562

and current reliability scores are used to infer the quality of the resource q_j, and the inferred quality $\hat{r}_j$ is then utilised to update the students' gained reliability score. However, there is an additional stage here where all other users linked to this set of students (ie, $(u_1, \dots, u_k)$) will receive an updated score from the most trustworthy moderator. As shown in Figure A1b, users' reliability would be updated by φ_i in this model based on the quality of their work and similarities to their peers s_{ik} who are directly connected to them due to their partnership in earlier moderations. For additional information, please refer to (Darvishi et al., 2021).

Text analysis

Length. The length of their comments can be used to estimate how much effort students put towards moderation. The $\mathrm{LC}_{N\times M}$ notation is introduced, where lc_{ij} represents the length of comments (ie, the number of words) provided by user u_i on resource q_j. Equation 2 is used to calculate the final rating:

$$\hat{r}_j = \frac{\sum_{i=1}^{k} \lambda_i \times d_{ij}}{\sum_{i=1}^{k} \lambda_i} \tag{2}$$

where λ_i is set to lc_{ij}, approximating u_i 'effort' in assessing q_j based on comment length. Informally, this strategy rewards students who provide a more detailed explanation for their assessment.

Relatedness. As shown in Equation 3, comment c_{ij} and resource q_j are encoded in a semantic vector space, and their cosine similarity in that space is measured to determine their relatedness.

$$\begin{gathered}\vec{c}_{ij} = \mathrm{Encoder}(c_{ij}) \quad \& \quad \vec{q}_j = \mathrm{Encoder}(q_j), \\ \mathrm{Relatedness}(c_{ij}, q_j) = \cos(\vec{c}_{ij}, \vec{q}_j)\end{gathered} \tag{3}$$

SBERT Reimers and Gurevych (2019) is utilised as the encoder function in Equation 3 to capture semantic relatedness rather than depending solely on exact lexical matching. Then, the cosine similarity score is used to reflect relatedness and ranges in [−1, 1], subsequently employed in Equation2 as the user decision weight λ_i.

Combined models

In these models, we explore integrating features from different inference models indicated above in λ_i of Equation 2. For instance, in a model consisting of relatedness from text analysis and trust from probabilistic models, λ_i would be a product of the relatedness of the submitted comment multiplied by the user's reliability score.

APPENDIX C

Dataset and Topics of comments in *Feedback on Review*

TABLE C1 An illustrative dataset used for *feedback on review* shows 14 students' feedback after receiving the peer review outcome on 4 resources

Resource ID	User ID	User role	Assessment outcome	User decision/ rating	Like	Dislike	User vote	User feedback on review
q1	u1	Assessor	Approve	Approve/3	1	0	Not Voted	–
q1	u9	Assessor	Approve	Approve/5	1	0	Disagree	Answer A is wrong, since logical data independence does not refer AT ALL to the physical structure of the data
q1	u10	Assessor	Approve	Approve/4	3	0	Not Voted	–
q1	u2	Author	Approve	NA/NA	0	0	Agree	Make it less wordy
q2	u3	Assessor	Approve	Approve/3	0	0	Agree	More feedback would be helpful
q2	u4	Assessor	Approve	Approve/4	1	0	Not Voted	–
q2	u5	Author	Approve	NA/NA	0	0	Agree	I would word the question a bit more differently as the moderators provided and improve the clarity of my answer
q3	u11	Assessor	Reject	Approve/3	0	0	Not Voted	–
q3	u12	Assessor	Reject	Reject/2	0	0	Agree	Some moderators could have provided more feedback to improve the presented analysis in my view
q3	u13	Assessor	Reject	Approve/3	2	0	Unsure	–
q3	u14	Author	Reject	NA/NA	0	0	Agree	Improving the entrepreneurial growth strategies and in-depth research to improve this work
q4	u6	Assessor	Approve	Approve/5	1	0	Disagree	I think that the resource had ‘outstanding’ content coverage as it has an accurately presented ER-diagram. There's nothing else left to be desired in this particular criterion for this resource
q4	u7	Assessor	Approve	Approve/4	2	0	Agree	I agree with this moderation
q4	u8	Author	Approve	NA/NA	0	0	Agree	I edited option B to make it clear that the user and the premium user are different people

TABLE C2 Topics of the additional comments provided by assessees and assessors in the feedback on reviews interface

	From			
	Assessees		Assessors	
Feedback on	Rate	Examples	Rate	Examples
Peers	18.4%	+ Thanks to my classmates give such useful and sincere feedback − Moderator 4 s assessment does not match with the others, and has dragged down my score, but their feedback and critique is not relevant to the resource I produced	39.8%	+ After reviewing my own notes, I agree with moderator 4 and would downvote my own moderation if I could − One of the moderators was far too harsh. This is a good question. Yes, the explanation could use some work but overall it's a good question and should not be denied"
Themselves	37.9%	+ I'm glad I wrote a good question − I was unaware that I should have included assumptions and entrepreneurial insights; thus, I would add these sections to improve my resource. I also believe that more research could have been undertaken to develop a deeper understanding which would have aided in the resources clarity	31.8%	+ I agree with my initial decision, the author may try to rephrase the answers to make it slightly different from the content on the slides − On reflection of the other moderators comments maybe my comments were too harsh as it appears they valued having more options
Outcome	3.4%	+ The review showed that the overall difficulty was good so I'm happy with the results but I could've used a different example − I edited the question fixing spelling mistakes and grammatical issues as advised by moderators. I am unsure why it did not pass when it has been edited	11.4%	+ Definitely worthy of a 4.3 and glad another user voted 5 as well. All moderators came to a common consensus that this question is well done − I do not think this resource should be accepted as it is basically all descriptive and offers no meaningful insight
System	2.3%	+ Through engaging in writing tasks, learners may have opportunities to review the knowledge they have learned to make progress rather than learning naturally. Thanks for the feedback. This provides me many thoughts − I think there is something wrong with the platform, a part of people did not saw the content	4.5%	+ Overall, the ripple platform was very helpful. I think the moderation process is very good it gives us the opportunity to understand which resource is good and which one is bad as a reader and as a author we can understand what could we do more so that the resource can be liked by everyone" − First years, myself included, aren't confident enough with the content to rate their own appraisal of content as high when giving criticism
Resource	90.8%	+ I have updated my question accordingly, thank you for pointing out the errors − I realise that the question and answers are incorrect and I would like to either edit or delete the question	76.1%	+ The best question I have encountered so far, the details and choice words are accurate and concise. Nice job! − This question requires no critical thinking, and even has an incorrect answer (D)

APPENDIX D

EXAMPLES OF METRICS AND DATASET USED FOR *INSTRUCTOR OVERSIGHT*

TABLE D1 An illustrative dataset used for *instructor oversight*, consisting of 27 resources spot-checked by 8 instructors, shows their flags' severity level (between [0,5])

Resource ID	Instructor ID	Outcome	Instructor action	Low effectiveness severity	Users reports severity	Assessors disagreement severity	Questionable distractors severity
q1	su1	Approve	Revised	0	0	4	0
q2	su1	Reject	Revised	0	2	0	0
q3	su2	Approve	Revised	0	0	2	0
q4	su2	Approve	Revised	0	0	4	0
q5	su3	Approve	Revised	0	2	0	3
q6	su3	Reject	Revised	0	0	0	3
q7	su4	Reject	Revised	1	0	0	3
q8	su5	Reject	Revised	0	0	0	5
q9	su5	Approve	Cleared	0	2	0	1
q10	su5	Reject	Cleared	0	2	0	0
q11	su5	Reject	Revised	0	2	0	0
q12	su5	Approve	Cleared	0	0	0	5
q13	su6	Reject	Revised	0	2	0	0
q14	su6	Approve	Revised	0	2	0	0
q15	su5	Reject	Revised	0	2	0	4
q16	su5	Approve	Cleared	0	0	0	3
q17	su7	Reject	Revised	5	2	0	1
q18	su6	Reject	Revised	0	2	0	2
q19	su6	Reject	Revised	3	5	0	4
q20	su6	Reject	Revised	5	2	0	0
q21	su6	Reject	Revised	5	2	0	0
q22	su6	Approve	Revised	0	2	3	0
q23	su6	Reject	Revised	0	2	2	5
q24	su8	Approve	Cleared	0	2	0	1
q25	su8	Reject	Revised	1	2	0	0
q26	su8	Reject	Revised	0	2	0	0
q27	su8	Reject	Revised	0	2	0	0

TABLE D2 Examples of metrics for spot-checking

Name	Description	Explainability	Example
Users' reports	Resources that have been reported as inappropriate or having inaccurate answers are prioritised based on the number of reports made by students	Provided comments of reporters	Resource has been reported 3 times **Report Reason** • Incorrect Answer. this relation should be in 2NF • Inappropriate Content. I believe the question is wrong first of all because the relation does not include E so in the answers it makes no sense when AE -> BC even though there is no E in the relation to begin with • Incorrect Answer. F in R should be E
Low effectiveness	Resources with a high proportion of downvotes (dislikes) compared to upvotes (likes), as determined by dividing the number of dislikes by the total votes (ie, Likes + Dislikes)	Comparing the effectiveness of the item to the course average	A high proportion of students have rated this resource as ineffective • Student Ratings: 17 • 3 Likes/14 Dislikes • Effectiveness: 17.6% • Course Average Effectiveness: 90%
Assessors' disagreement	Assessors' decision ratings on a resource diverge more than expected, which is prioritised by the standard deviation of their ratings	Showing the number of assessors and comparing their rating standard deviation with the course average	Disagreement between students is 2 times higher than average. There are significant discrepancies between the decisions on this resource • Student Assessors: 8 • Standard Deviation of Decision: 1.2 • Course Average Standard Deviation: 0.6
Questionable distractors	The author's answer is not the most popular, as indicated by the percentage of students' responses to incorrect distractors	Showing the response rate of incorrect answers and the number of students	Unexpected response distributions for 2 distractors. Student responses to 2 distractors in this resource are not in line with the solution given by the author based on 6 student responses Response Index **D.** Response Rate: 17% **F.** Response Rate: 33%