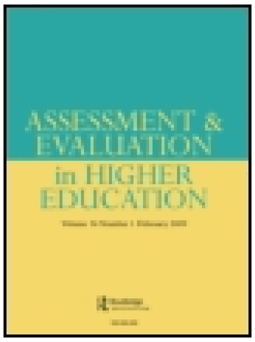


Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

The effects of rubrics on evaluative judgement: a randomised controlled experiment

George Gyamfi, Barbara E. Hanna & Hassan Khosravi

To cite this article: George Gyamfi, Barbara E. Hanna & Hassan Khosravi (2021): The effects of rubrics on evaluative judgement: a randomised controlled experiment, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2021.1887081

To link to this article: https://doi.org/10.1080/02602938.2021.1887081

Published online: 10 Mar 2021.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2021.1887081

The effects of rubrics on evaluative judgement: a randomised controlled experiment

George Gyamfi[a], Barbara E. Hanna[a] and Hassan Khosravi[b]

[a]School of Languages and Cultures, The University of Queensland, Saint Lucia, Australia; [b]Institute of Teaching and Learning Innovation, The University of Queensland, Saint Lucia, Australia

ABSTRACT

Rubrics have been suggested as a means to foster students' evaluative judgement, the capacity to appraise their own work and that of others; however, empirical evidence of rubrics' effectiveness is still emerging. This paper contributes findings from a randomised controlled experiment on the effect of rubrics on evaluative judgement. Participants were randomly assigned to one of two groups: a control group which evaluated peer-authored learning resources without the use of a rubric and an experiment group which carried out the evaluation using a three-item rubric based on (1) alignment with course content, (2) accuracy and (3) clarity. Both groups were asked to rate their confidence and provide comments to justify their scoring. The results showed a small effect size in increasing average agreement on the quality of learning resources in the experiment group. Analysis of comments reveals that criteria in and beyond the rubric guided participants' ratings of quality. The study provides evidence of the impact of rubrics on students' evaluative judgement and an example of how data-driven approaches and learning analytics can inform actionable design choices for embedding pedagogically supported strategies derived from the literature into actively operating educational technologies.

KEYWORDS

Evaluative judgement; rubrics; peer moderation; educational technologies; higher education

Introduction

Studies in higher education have emphasised the need to develop students' capability to understand and make judgements about the quality of work they and others produce, in any situation (Tai et al. 2018). The term evaluative judgement has emerged to describe this ability to evaluate quality in accordance with criteria or standards and to make appropriate judgements and improvements based on the appraisal (Boud 2000; Boud and Soler 2016; Tai et al. 2018; Panadero et al. 2019). This skill is considered necessary for any learning process and essential to all discipline-specific outcomes (Boud, Lawson, and Thompson 2015). The literature further suggests that once developed, this capability will not only give students the ability to apply a standard of quality to work but will also position them to use feedback effectively, to develop field-specific expertise and the autonomy to think critically, thus becoming reflexive and lifelong learners with knowledge of their evaluative potential (Nicol 2014; Boud and Soler 2016; Tai et al. 2016; Ajjawi et al. 2018).

CONTACT George Gyamfi g.gyamfi@uq.net.au

The recognition of the importance of evaluative judgement is often accompanied by questions on how it can be developed. While some studies present evaluative judgement as an innate capability waiting to be activated, others argue that students' evaluative expertise can be developed in a way similar to that of their content knowledge (Cowan 2010; Nicol 2014; Carless et al. 2018; Tai et al. 2018). Various strategies, such as the use of rubrics and criteria, self and peer assessment, exemplars, reviews and reflection have all been proposed to support the development of students' evaluative judgement (Boud, Lawson, and Thompson 2013; Nicol, Thomson, and Breslin 2014; Carless et al. 2018; Tai et al. 2018; Panadero et al. 2019). The potential role of rubrics in enhancing students' evaluative judgement has particularly been highlighted among these strategies (Boud, Lawson, and Thompson 2013; Ajjawi et al. 2018; Varela and Gregori-Giralt 2018; Tai et al. 2018).

While the efficacy of rubrics as instructional and assessment tools has been extensively studied in higher education literature, much of the research and commentary on the effects of rubrics to enhance students' evaluative judgement has been of a theoretical nature, with few direct examples of intervention studies or other kinds of empirical research (Boud, Lawson, and Thompson 2013; Bouwer et al. 2018). Moreover, there have been widespread concerns about the use of rubrics to make standards or criteria explicit to students. While some studies posit that rubrics are reliable tools for enhancing consistency in assessment, others have criticised their use as a means of conditioning students to comply with the stated standards or criteria without developing their autonomy (Torrance 2007). Other studies further argue that words, diagrams or symbols may not possess the necessary features to represent the criteria or standards and may lead to misunderstandings in application (Sadler 2014).

This study therefore aimed to empirically investigate the effects of rubrics on students' ability to evaluate the quality of peer-created learning resources. Specifically, we were interested in examining whether the use of a rubric would impact students' ratings of the quality of resources, help students reach a higher level of agreement on the quality of peer-authored resources, increase their confidence in their assessment of the learning resources and enable them better to articulate their judgements. To achieve our objective, the design of the study involved providing students with the opportunity to create and then evaluate the quality of learning resources. Participants were randomly assigned to one of two groups where the control group moderated the quality of peer created resources without a rubric and the experiment group carried out the same task with a three-item rubric based on (1) the alignment of a given resource with course content, (2) the accuracy of the resource and (3) the clarity of the resource. Participants from both groups then rated their confidence in their evaluation of their peers' work and provided feedback in the form of comments to justify their ratings. We analysed quantitative data from the ratings provided by participants and then examined the effects of the rubric on participants' confidence in their judgements and agreement on the quality of the peer-authored resources. Qualitative data in the form of comments were coded and thematically analysed.

Conceptual background

What is evaluative judgement?

The concept represented by the term evaluative judgement has been referred to in the literature on assessment for some decades now. Terms such as 'evaluative knowledge', 'evaluative expertise' (Sadler 1989), 'informed judgement' (Boud 2007) and 'judgement of usefulness' (Hastie and Dawes 2010) have all been used to designate the same construct. For example, in one of his earlier works on formative assessment, Sadler advocated for the engagement of students in assessment activities that aim at enhancing their 'evaluative knowledge' (1989). He argued that the involvement of students in authentic learning experiences leads to the development of their 'evaluative knowledge' through their understanding and appreciation of quality and how judgements are made. Students' understanding and appreciation of quality will in turn lead to the development not only of their evaluative capability, but also of their ability to self-monitor their learning processes. Subsequent studies have advanced the importance of involving students in assessment practices that allow them to exercise, apply and compare the quality of their work to the required standards (Boud 2000; Cowan 2010; Boud and Soler 2016).

Evaluative judgement integrates elements such as decision-making, self-regulated learning and meta-cognition into a core capability with a specific pedagogic focus on students' expertise in appraising work (Boud and Soler 2016; Ajjawi et al. 2018; Tai et al. 2018). For instance, in his work on sustainable assessment, Boud (2000) advances the importance of involving students in activities that aim to improve their capability to make such decisions, leading to the development of their skills for lifelong learning. Other studies have emphasised the need for students to have agency in their learning and assessment processes in order to develop their evaluative judgement. For example, Nicol (2010) criticised the over-dependence on one-way delivery of feedback in current assessment practices in higher education. He calls for the provision of opportunities that enable students to evaluate their learning with minimal guidance from instructors. Cowan (2010) posited that self and peer-assessment gives students agency over their learning and often results in the engagement of students in practical and transferable activities for the development of their judgemental expertise. Sadler (2010) argues in his later work that developing evaluative judgement is not only relevant for students' ability to make complex appraisals about work but also enables them to develop strategies to make further improvements of quality.

Central therefore to effective evaluative judgement is the development of students' expertise in identifying and understanding what constitutes good quality in their context or field of study. This capability enables students to become less dependent or reliant on instructors/others to determine the quality of their work, leading to the development of skills for lifelong learning (Ajjawi et al. 2018). However, despite evaluative judgement being recognised as a significant skill that needs attention to be improved, only a few recent studies have considered strategies to develop it (Nicol 2014; Barton et al. 2016; Tai et al. 2016; Carless et al. 2018; Panadero et al. 2019), and fewer still have been conducted to empirically verify the effects of the proposed strategies.

Principles for the development of evaluative judgement

Researchers posit that two principal and complementary components are essential for the development of students' evaluative judgement (Tai et al. 2018: Panadero et al. 2019). The first is an understanding of an appropriate standard of quality. Having such an understanding enables students to differentiate between work that meets the expected standard and work that does not. Standards of quality are contextual and are not conceptualised in the same way in different disciplines (Tai et al. 2018). Whereas some standards of quality may be written out, others exist in exemplars of work and experts and students may also have their own implicit understandings of quality which may be challenging to express (Ajjawi et al. 2018). The second component is the actual application of that understanding in the making of decisions regarding quality. This means that to nurture their evaluative judgement, students should be engaged in identifying appropriate criteria (either from an external source or through their own reflection) and in applying them to evaluate the quality of work of different standards.

Studies on the effectiveness of rubrics

Rubrics are instruments that guide evaluation by articulating the standards of quality expected of the object being assessed through a set of criteria, and may be used by instructors for formative and summative purposes and to provide feedback and grades (Popham 1997; Reddy and Andrade 2010; Brookhart 2013). Rubrics potentially enable instructors and students to determine whether a learning outcome has been achieved, what needs to be improved and where they must focus their efforts (Brookhart 2013). A number of empirical studies and reviews have been conducted on their effectiveness in just a little over a decade (Jonsson and Svingby 2007; Reddy and Andrade 2010; Panadero and Jonsson 2013; Brookhart and Chen 2015; Cockett and Jackson 2018; Panadero and Jonsson 2020). These studies provide evidence that rubrics are reliable tools for assessing performance, enhancing consistency in assessment and that co-creating rubrics with students enhances their acceptability and utility (Jonsson and Svingby 2007; Cockett and Jackson 2018).

A recent review of studies on arguments against the use of rubrics concludes that most of the evidence against the use of rubrics is based either on anecdotal or on personal experiences with little scientific value (Panadero and Jonsson 2020). Furthermore, these criticisms are mostly based on the summative use of rubrics with little emphasis on their formative function. In general, however, rubrics are considered effective tools that enhance students' learning, academic achievement, self-efficacy and self-regulation, if they are properly designed and adequately implemented (Panadero and Jonsson 2013; Brookhart and Chen 2015).

While the instructional and empirical value of rubrics for promoting assessment and students' learning is backed by extensive research, studies on their effectiveness in developing students' evaluative judgement have largely been theoretical. For instance, previous studies relating to rubrics and evaluative judgement posit that they make standards clear, communicate expert opinion to students and can be used to train them to develop their own expertise (Yuan et al. 2016; Tai et al. 2018): rubrics are said to assist students to develop, refine and agree on a standard of quality (Brookhart 2013; Jonsson 2014; Bearman et al. 2016; Tai et al. 2018; Varela and Gregori-Giralt 2018). These studies hypothesise that the involvement of students in evaluating work using rubrics leads to an understanding of standards of quality and the making of sound decisions about the quality of work in relation to those standards: rubrics guide students to make judgements of quality and accuracy, highlight procedures for assessing an object and function as a reference point that helps students justify their judgements.

This is not to say that all the work relating to evaluative judgement and rubrics has been speculative: there are some studies with direct examples of intervention or other kinds of empirical research. Of that limited existing work, one study compared marks given through students' self-assessment using a set of criteria with those of tutors to examine whether students' capacity to make judgements improved over time (Boud, Lawson, and Thompson 2013). The study concludes that rubrics were effective in enabling students to become better judges of their own work since their marks converged with those of tutors. The study however stresses that a one-time use of rubrics by students to make quality decisions does not necessarily mean that they can always make valid and informed judgements about every piece of work. Panadero and Romero (2014) examined the effect of rubrics on self-assessment. Their findings reveal that rubrics enhanced the accuracy of students' self-assessment scores and that engaging students in self-assessment without a rubric led to inaccuracies.

Varela and Gregori-Giralt (2018), in their study on the use of rubrics in developing arts students' professional judgement, concluded that involving students in the design of rubrics and moderating the quality of work with them enables the development of sound assessment skills. The study further inferred that the utilisation of rubrics as an instructional resource helps define standards and support the process of agreeing on a standard. These studies reveal both the potential of rubrics in improving students' learning and some limitations in their use. While they provide interesting insights into how rubrics can be utilised to enhance students' expertise, they do not examine the impact and effect-size of the use of rubrics on students' ability to judge the quality of resources. This was the intention of the current study.

Aim and research questions

The study aimed to empirically investigate the effects of rubrics on students' ability to evaluate the quality of peer-created learning resources. We hypothesised that rubrics would impact students' ratings of the quality of resources, increase agreement on the quality of resources, make students more confident in their assessment and enable them better to articulate their judgements of the quality of learning resources. The following questions guided the study:

1. Does the use of rubrics impact students' judgement of the quality of learning resources?
2. Does it increase agreement among students evaluating the same resource?
3. Does it affect students' confidence in their assessment of the resources?
4. How does it impact students' ability to articulate their judgement?

Method

Participants and setting

The study was conducted in an on-campus undergraduate course on Database Principles with 354 participants, at the University of Queensland. The course provided students with a basic understanding of concepts related to designing and implementing information systems necessary for advanced data management and analysis. Participants in the study were students who consented to be involved in the study and never withdrew that consent. The study collected data from the first five weeks of the semester during which students were assigned weekly assessment tasks of creating, moderating and engaging with learning resources (multiple choice questions, notes and worked examples) which would be shared with their peers. They moderated the peer-created learning resources to decide whether they should be released for use (Figure 1). In addition, participants gave feedback in the form of comments on each resource they moderated to justify their ratings. Resource authors were notified if their resource was approved, pending or denied. Authors of resources that were denied received the moderators' feedback which supported the improvement and re-submission of the resources. Figure 2 shows the interface used

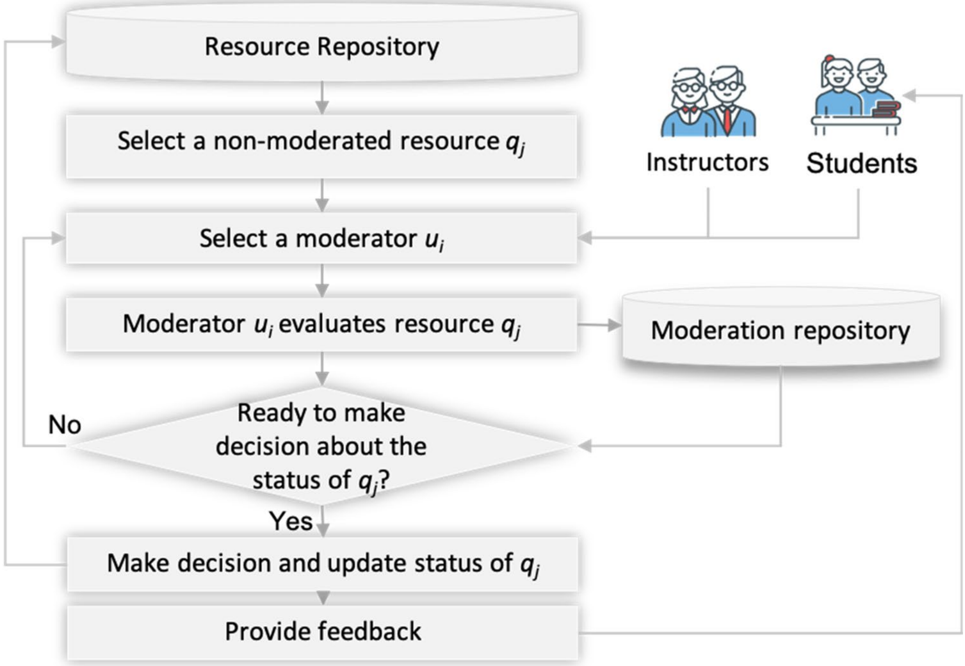
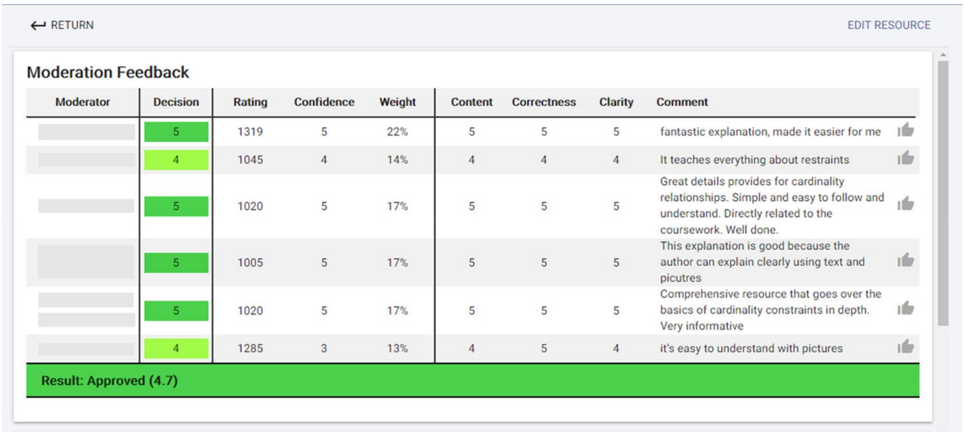

Resource Repository
Select a non-moderated resource q_j
Select a moderator u_i
Moderator u_i evaluates resource q_j
Moderation repository
Instructors
Students
No
Ready to make decision about the status of q_j?
Yes
Make decision and update status of q_j
Provide feedback

Figure 1. Overview of the moderation process in RiPPLE.



Moderation Feedback

Moderator	Decision	Rating	Confidence	Weight	Content	Correctness	Clarity	Comment
[redacted]	5	1319	5	22%	5	5	5	fantastic explanation, made it easier for me
[redacted]	4	1045	4	14%	4	4	4	It teaches everything about restraints
[redacted]	5	1020	5	17%	5	5	5	Great details provides for cardinality relationships. Simple and easy to follow and understand. Directly related to the coursework. Well done.
[redacted]	5	1005	5	17%	5	5	5	This explanation is good because the author can explain clearly using text and picutres
[redacted]	5	1020	5	17%	5	5	5	Comprehensive resource that goes over the basics of cardinality constraints in depth. Very informative
[redacted]	4	1285	3	13%	4	5	4	it's easy to understand with pictures

Result: Approved (4.7)

Figure 2. Feedback interface on RiPPLE.

for sharing this information. Resources that passed moderation were available for use by all students in the course and evaluated for their usefulness (a simple thumbs up or down vote).

Research tool: representation in peer personalized learning environment (RiPPLE)

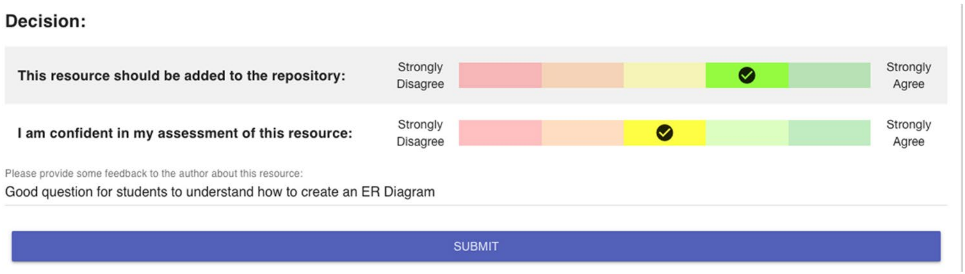
The study used an online learning platform called RiPPLE that employs learner centred and pedagogically informed approaches to engage students in an authentic learning experience (Khosravi et al. 2019). RiPPLE aims to enhance students' creativity and evaluative skills as experts-in-training by involving them in the development of a repository of high-quality learning resources. To do this, the platform provides students with a set of templates through which they can create a range of learning resources, namely multiple-choice questions, multi-answer questions, worked examples and an open-ended resource called 'notes'. Theoretically, both students and instructors can moderate learning resources; however, the system aims to minimise reliance on instructors. Therefore, RiPPLE depends on students by engaging them to moderate and judge the quality of the peer-created learning resources. Moderation may be supported by strategies such as the use of rubrics (of varying degrees of complexity: e.g. grading scheme/ open ended questions), self-assessment, peer moderation/assessment, peer feedback and provision of exemplars.

Figure 1 provides an overview of the moderation process in RiPPLE. From a repository of non-moderated resources, a resource *qj* is selected and assigned to an available student moderator *ui*. The moderator makes a judgement as to whether or not the resource should be included in the repository of course materials. RiPPLE then determines whether or not the resource needs to be evaluated by further moderators. At a high level of generality, RiPPLE seeks a minimum threshold of *k* moderations, where *k* by default is set to 3. RiPPLE then considers the level of agreement between moderators; if there is a strong agreement, then it will make a decision based on the formed consensus. Otherwise, it will request further moderations from students or an instructor, if available. RiPPLE works by using instructors' evaluation to make the final call whenever there is disagreement among student moderators. Once a decision has been made, RiPPLE updates the status of *qi* so that it is either approved and available to students for use or is denied and is unavailable to students.

RiPPLE enables researchers to conduct ethical, sound, large-scale, randomised, quasi-experimental and observational experiments. To date, RiPPLE has been used to support educational research in various fields including adaptive educational systems (AES), crowdsourcing, learner modelling, recommender systems, peer recommendation and dashboard visualisations (Abdi et al. 2019; Darvishi, Khosravi, and Sadiq 2020; Khosravi et al. 2020). Prior to the current study, the platform had not been used to investigate the effects of strategies that potentially enhance students' evaluative judgement through content creation, moderation and evaluation.

Experimental design

The study used a between-subject design where participants were randomly assigned to one of two groups of moderators. Participants were unaware of the group they were in to minimise biasing effect. The control group only rated, firstly, the quality of a resource with reference to whether it should be included in the learning repository and, secondly, their confidence in the accuracy of that rating. The experiment group completed a three-item rubric based on (1) alignment of the resource with course content, (2) the correctness of the resource and (3) clarity of the resource, before rating whether the resource should be included in the repository and indicating their confidence in their rating. The responses from both groups used a five-point Likert-scale where one represents strongly disagree, and five represents strongly agree. The participants finally provided feedback in the form of comments on each resource they moderated to justify their ratings. Figure 3 shows these two moderation interfaces (Figure 4). Each group moderated resources from a different pool of randomly allocated resources. This ensured that all moderations made for a specific resource were from the same group of students.



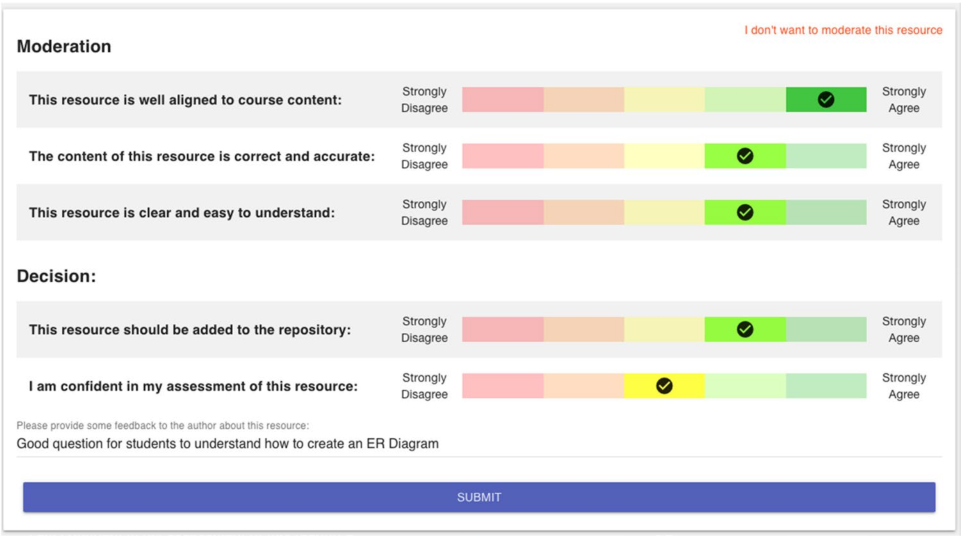
Decision:

This resource should be added to the repository: Strongly Disagree ○ ○ ○ ✓ ○ Strongly Agree

I am confident in my assessment of this resource: Strongly Disagree ○ ○ ✓ ○ ○ Strongly Agree

Please provide some feedback to the author about this resource:
Good question for students to understand how to create an ER Diagram

Figure 3. Moderation interface for control group.



Moderation

This resource is well aligned to course content: Strongly Disagree ○ ○ ○ ○ ✓ Strongly Agree

The content of this resource is correct and accurate: Strongly Disagree ○ ○ ○ ✓ ○ Strongly Agree

This resource is clear and easy to understand: Strongly Disagree ○ ○ ○ ✓ ○ Strongly Agree

Decision:

This resource should be added to the repository: Strongly Disagree ○ ○ ○ ✓ ○ Strongly Agree

I am confident in my assessment of this resource: Strongly Disagree ○ ○ ✓ ○ ○ Strongly Agree

Please provide some feedback to the author about this resource:
Good question for students to understand how to create an ER Diagram

Figure 4. Moderation interface for treatment group.

Data collection

During the 5-week period, the study collected quantitative data from participants' logs and qualitative data in the form of comments. Participant logs were their scoring/ratings of the overall quality of the peer-developed resources and their rating of their confidence in their assessment of the resources. Participants' overall quality ratings indicated whether a resource should be published for use by peers. Confidence ratings were interpreted as indications of participants' belief in their ability to exercise their evaluative judgement. Participants' logs were quantified, analysed and transformed into usable statistics. Their comments justifying their ratings were coded and thematically analysed. This mixed approach allowed for an in-depth investigation of the effects of the rubric on participants' ability to rate the quality of learning resources, from different perspectives.

Data analysis

A total of 2,212 moderations were completed. The control group ($n = 183$) and experiment group ($n = 171$) carried out 1,143 and 1,069 moderations respectively. We used a Mann-Whitney test to perform statistical analysis on the reported results.

Learner-sourced ratings

The quality ratings of peer created learning resources, which both groups of moderators supplied on a 5-point Likert scale, indicated whether a resource should be deposited for use in RiPPLE. For each group, the percentage, mean, standard deviation, median and p-value of participants' quality ratings were computed.

Confidence ratings

These were ratings of participants' confidence in their assessment of a resource. For each group, we collected and analysed the data in the form of percentages, mean, standard deviation, median and p-value.

Agreement

The standard deviation of the ratings for each resource was computed, where a smaller standard deviation represented a higher agreement. Using this information, the overall average (μ) and standard deviation (σ) of deviations across all the resources were then computed. This showed the extent to which the provision of a set of standards to students might reduce variation in quality judgements.

Length of comments

The average length of comments and the standard deviation were computed for each group to find out if use of a rubric reduced the need to comment.

Content of comments

The comments justifying participants' ratings are indicative of the criteria which constituted the students' understanding of appropriate standards. To analyse the comments, using the qualitative data analysis software package Nvivo, the entire data set from both groups was firstly read for familiarisation and to create provisional codes. The codes represented the criteria participants articulated in assessing the effectiveness or quality of the resources moderated. A total of 5% of the comments provided by each group was then manually coded to identify recurrent criteria

Table 1. Final code book generated from the manual coding of comments.

Codes	Definition	Examples
Accuracy	Comment refers to the correctness of the resource with respect to content.	**Positive:** 'Accurate resource. Great for revising main concepts of the relational model'. **Negative:** 'There is an error in the cardinality ratio'.
Alignment	Comment refers to the extent to which the resource matches content of the course.	**Positive:** 'Good question that tests general knowledge on the topic and checks if the student has actually paid attention during their classes'. **Negative:** 'I believe that the way the question is currently phrased does not completely rely on a course'.
Coherence	Comment refers to the level of consistency, logical connectedness and clarity of the resource.	**Positive:** 'This resource effectively links several concepts of the ER diagram'. **Negative:** 'The question is a bit ambiguous. From the question we don't know whether the relationship is from which set to the other set'.
Critical thinking	Comment refers to higher order learning or the development of critical thinking ability.	**Positive:** 'Very well thought through question. It requires students to think that step further'. **Negative:** 'This question helps understand the concept of referential integrity, but it doesn't require much thinking'.
Depth	Comment refers to the level of detail or complexity of the resource.	**Positive:** 'I also liked the complexity of the question especially the DOB not being in the format shown being used to throw off the user'. **Negative:** 'Elaborate more in the explanation on why it is more appropriate to create a new entity vs using what information has been given in the question'.
Difficulty	Comment refers to the level of difficulty or simplicity of the resource.	**Positive:** 'The difficulty level is intermediate and you gave the explanation which will be easy for people to understand their mistakes'. **Negative:** 'Personally, I think this question is a bit too easy. All of the options are just words mentioned during lecture so it would be easy to know that all of them are main concepts'.
Generic comments	General comments with little supporting detail.	**Positive:** 'The resource looks good'. **Negative:** 'Confusing'
Language	Comment refers to linguistic accuracy of the resource.	**Negative 1**: 'Grammar needs work here. "description" should be "descriptions," "identify" should be "identifies"'. **Negative 2:** 'There is a spelling mistake in the answers'.
Practical relevance	Comment refers to practicality, applicability or relevance of the resource to real life or other course settings.	**Positive:** 'The explanation improves the understanding of DBMSs in real life'. **Negative:** 'Agree. But consider in real life, people would think this in a perspective of human and car. Human without car still exist, car without human will be garbage'.
Resource type	Comment refers to the type of resource. e.g. distractor effectiveness (MCQ), note, worked example	**Worked example:** 'The example is quite interesting, but the relationships between different entities in the question aren't very clear and the worked example seem to come out of nowhere'. **MCQ:** 'This is because most multiple-choice questions do not have an 'all of the above' answer, so it is good for the user's engagement with the resource to see a different type of answer option'.

that were applied. The list of manually generated codes was defined as part of a code book which would be applied to a wider sample (see Table 1). This allowed for simplification and focus on specific aspects of the data.

A total of 515 and 481 comments representing 45% of the total comments provided were randomly sampled from the control and experimental group respectively for the final analysis. A total of 557 and 515 tags (instances of codes) were applied to the comments from the control and the experimental group respectively. For each group, the frequency of each code was computed using the matrix coding function in Nvivo. Based on this, the percentage of each code out of the total tags applied to the randomly sampled comments from each group was computed. This allowed for an analysis of the rate at which a particular criterion was referenced in the assessment of quality.

To ensure the reliability of the coding, 30% of the randomly sampled comments from each group, 155 comments from the control group and 144 comments from the experiment group, were double coded by an independent researcher. The double coder was orientated on how to apply the codes as stipulated in the code book and any misunderstandings about the definition of a criterion were clarified. The inter-rater agreement was excellent, at 96.8% across all codes. In addition, the overall Cohen's kappa coefficient across all the codes was 0.93; this coefficient measured the degree of inter-rater agreement in applying the codes to the comments and accounts for the probability that the coders guessed or applied some of the codes due to uncertainty or by chance.

Ethical considerations

While randomised controlled experiments remain the gold standard test for establishing causality in education research, there have been debates about how, without due care, they may disadvantage students in one of the experimental groups by providing poorer learning opportunities (Morrison 2001). However, in the present study, the possible advantage for one group of students related to the development of their evaluative judgement, not to their mastery of course content and their grades. All students, irrespective of their participation in the study or the group of moderators to which they were assigned, had access to the approved learning resources. The following processes were also followed to ensure compliance with principles of good research: 1. students were first presented with a consent form to seek their permission to run educational experiments using their data; 2. students could withdraw their consent at any time with no negative consequences, and their data were removed from the study. 3. the collected data were fully anonymised and were only accessible to the researchers.

Results

Does the use of rubrics impact students' judgement of the quality of learning resources?

To answer this question, participants' ratings of the quality of resources were analysed. The results reported in Table 2 reveal that, in general, both the treatment group and control group provided positive ratings (agree to strongly agree) indicating that the resources should be included in the pool of available resources; however, the ratings from the treatment group tended to be higher (μ = 4.08, $Mdn = 4$, $\sigma = 1.06$) compared to the control group (μ = 3.94, $Mdn = 4$, $\sigma = 1.00$);$U = 55038$, p *<.01*. This shows that the rubric made a difference since the experimental group gave slightly higher ratings than the control group. One possible explanation for the provision of higher average ratings is that the rubric provided was perhaps simpler and more generous than the implicit rubric the control group had in mind when moderating. The analysis of the comments will allow us to see to some extent whether the rubric was indeed simpler than the criteria used by the students.

Does the use of rubrics increase agreement among students evaluating the same resource?

The analysis of average agreement suggests that the rubric aided participants in the treatment group (μ = 0.80, $Mdn = 0.81$, $\sigma = 0.43$) to achieve a slightly higher agreement on their quality judgements compared to participants in the control group (μ = 0.85,

Table 2. Analysis of data on rating and confidence.

		Control %	Treatment %	U and P values
Ratings	Strongly Agree	34.00%	44.3%	$U = 55038$, $p < 0.001$
	Agree	37.80%	31.62%	
	Neutral	18.70%	14.69%	
	Disagree	7.30%	5.80%	
	Strongly Disagree	2.20%	3.46%	
Confidence	Strongly Agree	43.10%	54.91%	$U = 530214$, $p < 0.001$
	Agree	40.30%	34.33%	
	Neutral	12.60%	8.79%	
	Disagree	3.00%	1.40%	
	Strongly Disagree	1.00%	0.56%	

Table 3. Analysis of data on agreement and length of comments.

	Control	**Treatment**	**U and P**
	μ±σ	**μ±σ**	**values**
Agreement	0.85 ± 0.36	0.80 ± 0.43	$U = 18470$, $p = 0.271$
Length of comments	18. ± 2 1.94	17.9 ± 22.5	$U = 591332$, $p = 0.191$

Mdn = 0.82, $\sigma = 0.36$) (Table 3). However, the difference between the two groups was not statistically significant; $U = 18470$, $p = 0.271$. An effect size analysis using Cohen's d showed a small effect of the rubric on students' agreement *d = 0.13*. We speculate that the rubric missed references to some criteria the participants associated with quality and that the use of a more comprehensive rubric which aligned with students' implicit expectations of an effective resource may lead to a higher agreement among moderators.

Does the use of rubrics affect students' confidence in their assessment of the resources?

Students' ratings of confidence in their assessment of the peer created resources were analysed to provide statistical information on the average, standard deviation and the p value for each group. The results indicate that both groups of students demonstrated high confidence in their ratings (agree to strongly agree); however, compared to students in the control group (μ = 4.22, *Mdn* = 4, σ =.80) students in the treatment group (μ = 4.42, *Mdn* = 5, σ =.88) showed a statistically significant higher level of confidence in their ratings; $U = 530214$, p <.01. We speculate that the availability of an explicit standard of quality, provided by the rubric, contributed to a higher level of confidence by providing validation of the criteria used.

How does the use of rubrics impact students' ability to articulate their judgement?

Findings from the length and contents of comments provided by each group were analysed to provide answers to this question.

Length of comments

With regards to the length of comments, measured in words, the control group (μ = 18.2, *Mdn* = 12, $\sigma = 19.4$) provided slightly longer comments than the treatment group (μ = 17.9, *Mdn* = 11, *σ* = 22.5), however the difference between these two results was not statistically significant; *U* = 591332.5, *p* = 0.191 (Table 3). This suggests that the students did not feel that using a rubric removed the necessity to provide feedback.

Content of comments

An analysis of the comments reveals that both groups were able to articulate their perceptions of the quality of the resources they moderated. This is evident in the criteria expressed in the comments provided. Among these were some criteria that were already captured by the rubric and others that were not referenced in it. We report findings on the percentages of codes applied to comments from each group.

Criteria captured by the rubric

Both groups used criteria which were captured in the rubric to evaluate the quality of the resources (see Figure 5a). Specifically, the participants were concerned with the accuracy of resources, alignment with the course content and coherence/clarity of the resources. While we might expect that the experiment group, with access to the rubric, would express their judgements of quality based on criteria expressed in it, it is noteworthy that the control group also articulated the criteria in the rubric even though they had no access to it. But were these criteria equally important to the two groups?

The analysis revealed that 'accuracy of the content' constituted 11.49% of the codes applied to comments from the control group compared to 9.71% for the treatment group. That the control group made more references to this quality perhaps indicates that accuracy is a fairly obvious criterion to be applied to learning resources, salient even to students without the rubric. Both groups were interested in the correctness of the resources and their accuracy with respect to the content of the course. There was a difference between the percentage of codes concerning 'alignment', the extent to which the resource reflected the content of the course: 8.62% for the control group compared to 16.90% for the experiment group. This contrast suggests that the inclusion of this criterion in the rubric made the experiment group more aware of its relevance. The participants were further concerned with 'coherence and clarity', a criterion that was partially captured in the rubric under 'clear and easy to understand'. However, the participants were not only interested in the clarity of the resources, but also their consistency and logical connectedness (Table 1). This made up 20% of the codes applied to comments from the experiment group compared to 17.41% for the control group.

Overall, both groups made references to all the criteria that were captured in the rubric, but the experiment group was more aware of them. This signifies that these criteria were of value to students in their assessment of the quality of learning resources.

Criteria not captured by the rubric

The analysis further revealed that the participants expressed some criteria that were not captured in the rubric, namely: potential for critical thinking, depth, difficulty, practical relevance and features specific to a resource type (see Figure 5b). The criterion 'depth' made up 13.11% of the codes applied to comments provided by the control group and 9.90% of codes applied to comments from the experimental group. This suggests that the participants were concerned about the level of detail of the information provided in the resources.

The tag 'difficulty' constituted 19.03% of the codes applied to comments provided by the experimental group compared to 17.77% for the control group. As regards difficulty, the comments were both positive and negative. Firstly, some attested that particular resources were perceived to be difficult, but at the right level to enhance learning of the course content. Other comments indicated that specific resources were considered overly complicated, extremely difficult and would not encourage learning. On the other hand, some resources were seen as simple but effective, while other simple resources were judged ineffective for learning. For

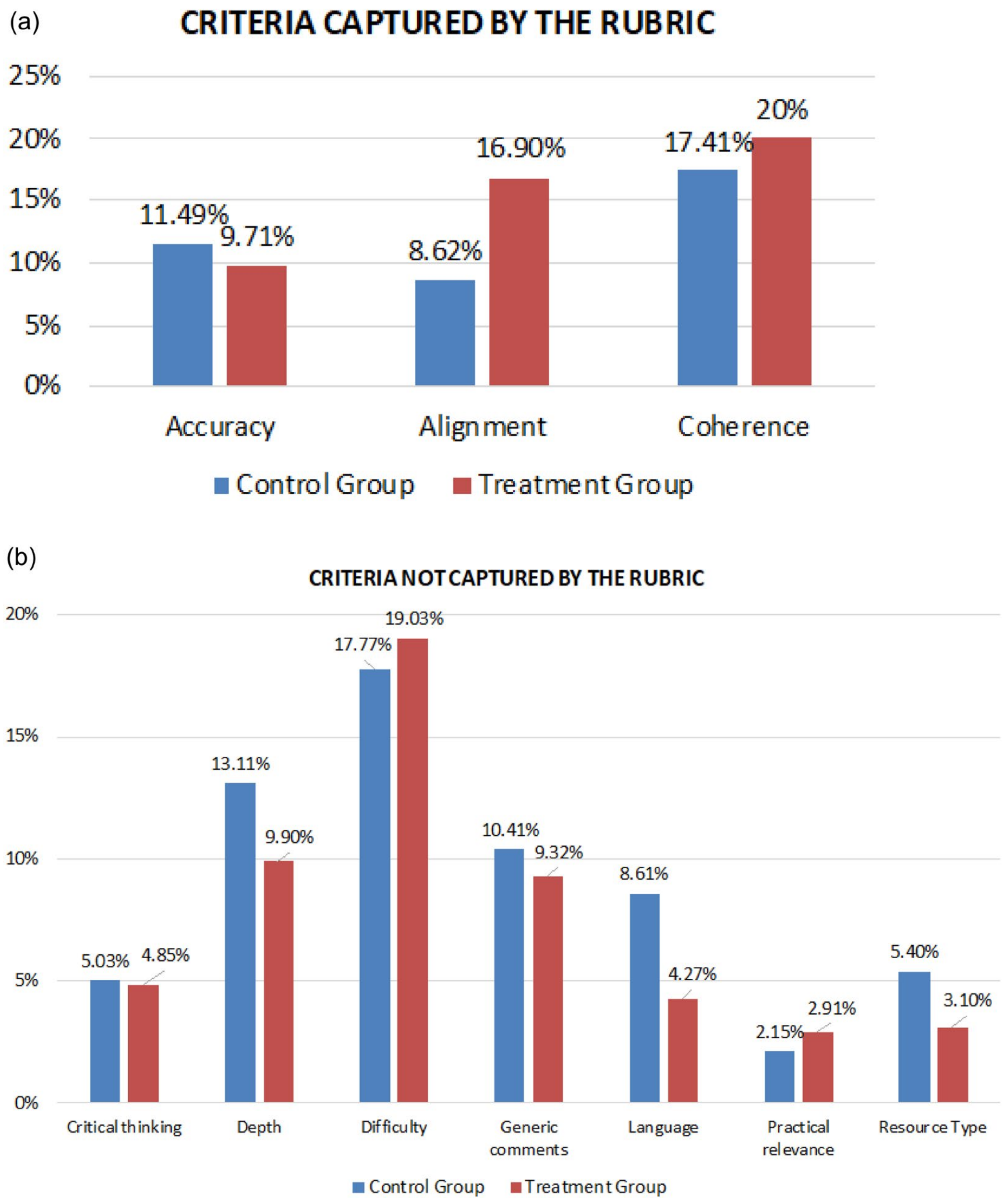


Figure 5. The percentage of codes applied to comments from each group.

resources that were deemed simple and ineffective, the student moderators suggested increasing the difficulty level in order to improve learning.

In relation to 'accuracy of language', the difference between the percentage of codes, 8.61% for the control group compared to 4.27% for the experimental group, suggests that the control group paid attention to features that did not relate to the actual content of a learning resource, whereas the rubric focused attention on the potential for learning. 'Generic comment' made up 10.41% of the codes applied to comments provided by the control group compared to 9.32% for the experimental group. The provision of more 'generic comments' by the control group could be attributed to the absence of a rubric to guide and support the articulation of their judgement.

In addition, further criteria appeared in the comments, although to a lesser extent. These include the potential of a resource to enhance critical thinking, its practical relevance and features specific to a resource type. There was a minimal difference between the percentages of codes concerning 'critical thinking', 5.03% for the control group and 4.85% for the experiment group. 'Practical relevance' comprised 2.15% of the codes applied to comments provided by the control and 2.91% of codes applied comments from the experiment group. In relation to features specific to a resource type, this was more important to the control group (5.40% of tags), which was concerned about the effectiveness of the distractors in the multiple-choice questions, compared to the experiment group (3.10% of tags). Even though these criteria were less referenced in both groups they reveal students' understanding of what constitutes quality and show evidence of their thought processes and justifications of their ratings.

Overall, the analysis of comments reveals that the participants referred both to criteria that were captured by the rubric and criteria that were not. The use of these criteria independent of the rubric indicates that students have an implicit understanding of what constitutes quality in relation to learning resources and were able to apply it in practice.

Difference in commenting behavior

Further analysis was conducted to determine whether the criteria each group applied to assess the effectiveness of the resources they moderated were significantly different. This analysis aimed at testing the null hypothesis that there was no significant difference in the criteria participants used. To do this, the percentages of each code applied to comments provided by each group were subjected to a chi-square goodness of fit test. The results showed a non-statistically significant difference between the criteria the participants applied in assessing the effectiveness of the resources across all the codes except for alignment and language, p <.05.

Discussion

The study aimed at investigating the effects of rubrics on students' ability to evaluate the quality of learning resources. It revealed that the use of rubrics can positively but slightly impact students' agreement in assessing the quality of learning resources. The differences between the ratings from the experimental and control groups imply that rubrics can influence how students attend to quality. It is therefore essential that rubrics provide a useful set of criteria that reflect quality in students' field of study to enable them to practise the application of relevant standards (Tai et al. 2018).

In addition, the rubric led to greater agreement in judging the quality of the resources in the experiment group compared to the control group. This is an indication that the rubric impacted participants' quality judgements by enabling them to be consistent across the group. This finding aligns with a study by Panadero and Romero (2014) which concludes that rubrics can contribute to the consistency and accuracy of students' scoring. However, the difference in level of agreement was not statistically significant in our experiment. The high level of intragroup agreement and small difference between the groups suggest that the concept of quality held by students from the control group was similar to that of the rubric.

Furthermore, the study reveals a difference in the level of confidence in participants' assessment of the peer-authored resources, with the experimental group being more confident. This presumably suggests that the expected standard of quality was stated clearly in the rubric which they used, and that it could be applied without confusion. Though the findings show that students' confidence in making quality judgements was high even without the use of rubric, the difference in ratings suggests that the rubric served as a useful tool to scaffold students' decision-making, giving them confidence in it. This finding is in line with studies that show that rubrics provide transparency which may reduce anxiety and give students confidence in the assessment process (Panadero, Tapia, and Huertas 2012).

The provision of comments by the experiment group, even though they had access to the rubric which already articulated and justified their ratings, indicates that using a rubric did not eliminate the need to give comments. The provision of comments by both groups suggests further that students are able to give feedback on the quality of resources whether or not they are guided by rubrics. There were however some differences between the groups. The analysis shows that in general, the control group gave slightly longer comments than the experiment group. The absence of a rubric to justify their decision and the need to provide enough rationale to support their judgements could account for this. In addition, the control group lacked specifically defined criteria with key descriptors of quality to guide their judgements. This might have led them to provide both more generic comments (short, with no justification) and slightly longer comments (with fuller justifications) than the experimental group. This attests to the influence of the rubric on how students attend to quality.

The findings further show that the two groups articulated a similar set of criteria in making decisions about the quality of the resources they evaluated, applying both rubric and non-rubric criteria. A possible explanation for the similarity in the criteria used is that the participants had a similar implicit understanding of quality in their field. It therefore appears that students' evaluative judgement or sense of quality has already developed to some extent in relation to learning resources in their field. A further implication is that students are likely to turn to additional criteria which are already of relevance to them if they perceive that the given criteria (here, those supplied in the rubric) do not enable them to fully evaluate essential features of a resource. The use of additional criteria by the treatment group supports Sadler's observation that words, diagrams or symbols may not possess the necessary features to represent the criteria or standards (2014).

In sum, the ability of students to not only rate the quality of the resources but also give comments to justify their ratings demonstrates their application of evaluative judgement. Both groups made comments which led to the confirmation of the criteria captured on the rubric as indicators of quality and to the discovery of additional criteria that influenced their judgements. Their application of an implicit set of criteria shows their understanding of standards in their field and a further demonstration of their evaluative judgement, their ability to make decisions about quality and justify them.

Conclusion and future work

The study demonstrates the effects of rubrics on the evaluation of the quality of learning resources, as part of a meaningful goal-oriented activity, the creation of a shared repository of online learning activities. Our results provide evidence that rubrics have a positive but slight impact on students' ability to make judgements in that there was a higher level of agreement but a small effect size of the rubric on the ratings of the quality of learning resources. The explicit statement of quality provided by the rubric was relatively consistently applied. Students operating without such guidance also used a common understanding of quality. That the difference in level of agreement was not statistically significant shows that, even without rubrics, students are capable of judging the quality of resources in a coherent, meaningful way. Students used a similar set of criteria for judgement based on their implicit understanding of quality whether they are provided with a rubric or not. While the average length of comments was shorter in the treatment group, students without the guidance of a rubric were more likely to give very short generic comments compared to students guided by a rubric to justify their decisions.

The question remains however as to whether these differences would be sustained, would attenuate, or become more marked over time. The data set utilised in this study was from the first five weeks when the rubric was newly introduced to students. The study continues to investigate whether the effects of rubrics identified here were maintained throughout the rest of the course. A further question relating to time is that of time-on-task and whether the use of a rubric adds to students' processing time or makes it more efficient.

Despite these limitations, this experiment proved useful in revealing the standards students apply in making judgements about the quality of learning resources. The analysis of the comments led to the discovery of additional criteria that influenced participants' ratings and which can therefore be seen to be relevant to their context. An improved rubric which incorporates some of these criteria will be designed for use in subsequent offerings of the course. Further studies will aim to replicate this study across different cohorts using the improved rubrics with a larger number of students.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

George Gyamfi is a PhD candidate in the School of Languages and Cultures at The University of Queensland. George aims to apply theories and concepts from the field of education in a novel and innovative way combined with insights from Information and Communication Technologies to make practical contributions to learning and research in higher education. His current research aims at investigating the effects of strategies for the development of students' evaluative judgement using an adaptive online learning system.

Dr Barbara E. Hanna is Senior Lecturer in French in the School of Languages and Cultures, The University of Queensland. She has worked on applications of Information and Communication Technologies to language and culture teaching; and on learner identity and agency.

Dr Hassan Khosravi is a Senior Lecturer In Data Science and Learning Analytics at The University of Queensland. In his research, he draws on theoretical insights driven from the learning sciences and human-computer interaction and exemplary techniques from the fields of machine learning and educational data mining to design, implement, validate and deliver socio-technical systems that improve student learning.

ORCID

George Gyamfi http://orcid.org/0000-0002-2589-1706
Barbara E. Hanna http://orcid.org/0000-0002-8450-4243
Hassan Khosravi http://orcid.org/0000-0001-8664-6117

References

Abdi, S., H. Khosravi, S. Sadiq, and D. Gasevic. 2019. "A Multivariate ELO-Based Learner Model for Adaptive Educational Systems." Proceedings of the 12th International Conference on Educational Data Mining (pp. 228–233). Montreal, Canada.

Ajjawi, R., J. Tai, P. Dawson, and D. Boud. 2018. "Conceptualising Evaluative Judgement for Sustainable Assessment in Higher Education." In *Developing Evaluative Judgement in Higher Education: Assessment for Knowing and Producing Quality Work*, edited by D. Boud, R. Ajjawi, P. Dawson, and J. Tai Hong-Meng, pp.7-17. Abingdon: Taylor & Francis.

Barton, K. L., S. J. Susie, S. McAleer, and R. Ajjawi. 2016. "Translating Evidence-Based Guidelines to Improve Feedback Practices: The interACT Case Study." *BMC Medical Education* 16 (1): 1–12. doi:10.1186/s12909-016-0562-z.

Bearman, M., P. Dawson, D. Boud, S. Bennett, M. Hall, and E. Molloy. 2016. "Support for Assessment Practice: Developing the Assessment Design Decisions Framework." *Teaching in Higher Education* 21 (5): 545–556. doi:10.1080/13562517.2016.1160217.

Boud, D. 2000. “Sustainable Assessment: Rethinking Assessment for the Learning Society.” *Studies in Continuing Education* 22 (2): 151–167. doi:10.1080/713695728.

Boud, D. 2007. “Reframing Assessment as If Learning Were Important.” In *Rethinking Assessment in Higher Education*, edited by D. Boud, and N. Falchikov, pp.14–26. London: Taylor & Francis.

Boud, D., R. Lawson, and D. G. Thompson. 2013. “Does Student Engagement in Self-Assessment Calibrate Their Judgement over Time?” *Assessment & Evaluation in Higher Education* 38 (8): 941–956. doi:10.1080/02602938.2013.769198.

Boud, D., R. Lawson, and D. G. Thompson. 2015. “The Calibration of Student Judgement through Self-Assessment: Disruptive Effects of Assessment Patterns.” *Higher Education Research & Development* 34 (1): 45–59. doi:10.1080/07294360.2014.934328.

Boud, D., and R. Soler. 2016. “Sustainable Assessment Revisited.” *Assessment & Evaluation in Higher Education* 41 (3): 400–413. doi:10.1080/02602938.2015.1018133.

Bouwer, R., L. Marije, B. Pieterjan, and D. M. Sven. 2018. “Applying Criteria to Examples or Learning by Comparison: Effects on Students’ Evaluative Judgment and Performance in Writing.” In *Frontiers in Education* 3: 86. doi:10.3389/feduc.2018.00086.

Brookhart, S. M. 2013. *How to Create and Use Rubrics for Formative Assessment and Grading*. Alexandria, Virginia: ASCD.

Brookhart, S. M., and F. Chen. 2015. “The Quality and Effectiveness of Descriptive Rubrics.” *Educational Review* 67 (3): 343–368. doi:10.1080/00131911.2014.929565.

Carless, D., K. K. H. Chan, J. To, M. Lo, and E. Barrett. 2018. “Developing Students’ Capacities for Evaluative Judgement through Analysing Exemplars.” In *Developing Evaluative Judgement in Higher Education: Assessment for knowing and producing quality work,* edited by D. Boud, R. Ajjawi, P. Dawson & J. Tai, 108–116. London: Routledge.

Cockett, A., and C. Jackson. 2018. “The Use of Assessment Rubrics to Enhance Feedback in Higher Education: An Integrative Literature Review.” *Nurse Education Today* 69: 8–13. doi:10.1016/j.nedt.2018.06.022.

Cowan, J. 2010. “Developing the Ability for Making Evaluative Judgements.” *Teaching in Higher Education* 15 (3): 323–334. doi:10.1080/13562510903560036.

Darvishi, A., H. Khosravi, and S. Sadiq. 2020. “Utilising Learnersourcing to Inform Design Loop Adaptivity.” In *Addressing Global Challenges and Quality Education, EC-TEL 2020, Lecture Notes in Computer Science*, Vol. 12315, edited by C. Alario-Hoyos, M.J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, S.M. Dennerlein, pp. 332-346. Cham: Springer. doi:10.1007/978-3-030-57717-9_24.

Hastie, R., and R. Dawes. 2010. *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Los Angeles: Sage.

Jonsson, A. 2014. “Rubrics as a Way of Providing Transparency in Assessment.” *Assessment & Evaluation in Higher Education* 39 (7): 840–852. doi:10.1080/02602938.2013.875117.

Jonsson, A., and G. Svingby. 2007. “The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences.” *Educational Research Review* 2 (2): 130–144. doi:10.1016/j.edurev.2007.05.002.

Khosravi, H., K. Kitto, and J. J. Williams. 2019. “RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities.” *Journal of Learning Analytics* 6 (3): 91–105.

Khosravi, H., G. Gyamfi, B. E. Hanna, and J. Lodge. 2020. “Fostering and Supporting Empirical Research on Evaluative Judgement via a Crowdsourced Adaptive Learning System.” In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 83–88. Frankfurt, Germany. doi: 10.1145/3375462.3375532

Morrison, K. 2001. “Randomised Controlled Trials for Evidence-Based Education: Some Problems in Judging What Works.” *Evaluation & Research in Education* 15 (2): 69–83. doi:10.1080/09500790108666984.

Nicol, D. 2010. “The Foundation for Graduate Attributes: Developing Self-Regulation through Self and Peer Assessment.” *The Quality Assurance Agency for Higher Education* [Scotland]. http://tinyurl.com/avp527r

Nicol, D. 2014. “Guiding Principles for Peer Review: Unlocking Learners’ Evaluative Skills.” In *Advances and Innovations in University Assessment and Feedback*, edited by C. Kreber, C. Anderson, J. McArthur, & N. Entwistle,197–224. Edinburgh.

Nicol, D., A. Thomson, and C. Breslin. 2014. “Rethinking Feedback Practices in Higher Education: A Peer Review Perspective.” *Assessment & Evaluation in Higher Education* 39 (1): 102–122. doi:10.1080/02602938.2013.795518.

Panadero, E., J. Broadbent, D. Boud, and J. M. Lodge. 2019. “Using Formative Assessment to Influence Self- and Co-Regulated Learning: The Role of Evaluative Judgement.” *European Journal of Psychology of Education* 34 (3): 535–557. doi:10.1007/s10212-018-0407-8.

Panadero, E., and A. Jonsson. 2013. “The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review.” *Educational Research Review* 9: 129–144. doi:10.1016/j.edurev.2013.01.002.

Panadero, E., and A. Jonsson. 2020. “A Critical Review of the Arguments against the Use of Rubrics.” *Educational Research Review* 30: 100329. doi:10.1016/j.edurev.2020.100329.

Panadero, E., and M. Romero. 2014. “To Rubric or Not to Rubric? The Effects of Self-Assessment on Self-Regulation, Performance and Self-Efficacy.” *Assessment in Education: Principles Policy & Practice* 21 (2): 133–148. doi:10.1080/0969594X.2013.877872.

Panadero, E., J. A. Tapia, and J. A. Huertas. 2012. “Rubrics and Self- Assessment Scripts Effects on Self-Regulation, Learning and Self-Efficacy in Secondary Education.” *Learning and Individual Differences* 22 (6): 806–813. doi:10.1016/j.lindif.2012.04.007.

Popham, W. J. 1997. “What’s Wrong–and What’s Right–with Rubrics.” *Educational Leadership* 55 (2): 72–75.

Reddy, Y. M., and H. Andrade. 2010. “A Review of Rubric Use in Higher Education.” *Assessment & Evaluation in Higher Education* 35 (4): 435–448. doi:10.1080/02602930902862859.

Sadler, D. R. 1989. “Formative Assessment and the Design of Instructional Systems.” *Instructional Science* 18 (2): 119–144. doi:10.1007/BF00117714.

Sadler, D. R. 2010. “Beyond Feedback: Developing Student Capability in Complex Appraisal.” *Assessment & Evaluation in Higher Education* 35 (5): 535–550. doi:10.1080/02602930903541015.

Sadler, D. R. 2014. “The Futility of Attempting to Codify Academic Achievement Standards.” *Higher Education* 67 (3): 273–288.

Tai, J., R. Ajjawi, D. Boud, P. Dawson, and E. Panadero. 2018. “Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work.” *Higher Education* 76 (3): 467–481. doi:10.1007/s10734-017-0220-3.

Tai, J. H. M., B. J. Canny, T. P. Haines, and E. K. Molloy. 2016. “The Role of Peer-Assisted Learning in Building Evaluative Judgement: Opportunities in Clinical Medical Education.” *Advances in Health Sciences Education* 21 (3): 659–676. doi:10.1007/s10459-015-9659-0.

Torrance, H. 2007. “Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-Secondary Education and Training Can Come to Dominate Learning.” *Assessment in Education* 14 (3): 281–294. doi:10.1080/09695940701591867.

Varela, J. L. M., and E. G. Gregori-Giralt. 2018. “Rubrics for Developing Students’ Professional Judgement: A Study of Sustainable Assessment in Arts Education.” *Studies in Educational Evaluation* 58: 70–79.

Yuan, A., K. Luther, M. Krause, S. I. Vennix, S. P. Dow, and B. Hartmann. 2016. “Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques.” In Proceedings of the 19th ACM Conference on Computer- Supported Cooperative Work & Social Computing, 1005–1017. San Francisco, CA.