JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 8(3), 117—132. https://doi.org/10.18608/jla.2021.7206

Bridging the Gap Between Theory and Empirical Research in Evaluative Judgment

Hassan Khosravi[1], George Gyamfi[2], Barbara E. Hanna[3], Jason Lodge[4], Solmaz Abdi[5]

Abstract

The value of students developing the capacity to accurately judge the quality of their work and that of others has been widely studied and recognized in higher education literature. To date, much of the research and commentary on evaluative judgment has been theoretical and speculative in nature, focusing on perceived benefits and proposing strategies seen to hold the potential to foster evaluative judgment. The efficacy of the strategies remains largely untested. The rise of educational tools and technologies that generate data on learning activities at an unprecedented scale, alongside insights from the learning sciences and learning analytics communities, provides new opportunities for fostering and supporting empirical research on evaluative judgment. Accordingly, this paper offers a conceptual framework and an instantiation of that framework in the form of an educational tool called RiPPLE for data-driven approaches to investigating the enhancement of evaluative judgment. Two case studies, demonstrating how RiPPLE can foster and support empirical research on evaluative judgment, are presented.

Notes for Practice

- This paper presents a conceptual framework for the development of educational tools that would promote both the development of evaluative judgment and research into it, along with a referenced implementation of that framework.
- The framework provides explicit guidance to educational technology developers and instructors through a co-design process to facilitate the development of tools and metrics to ascertain the impact of pedagogically supported strategies for developing students' evaluative judgment.
- In order to support the development of evaluative judgment and research into it, evaluative judgment strategies and learning analytics/metrics to verify their effect need to be integrated into educational tools and learning design from the outset.

Keywords

Evaluative judgment, educational technologies, learnersourcing, empirical educational research

Submitted: 15/05/2020 — **Accepted:** 07/04/2021 — **Published:** 03/11/2021

Corresponding author [1] *Email: h.khosravi@uq.edu.au Address: Institute for Teaching and Learning Innovation, The University of Queensland, St Lucia QLD 4072, Australia. ORCID ID: https://orchid.org/0000-0001-8664-6117*

[2] *Email: g.gyamfi@uq.net.au Address: School of Languages and Cultures, The University of Queensland, St Lucia QLD 4072, Australia. ORCID ID: https://orchid.org/0000-0002-2589-1706*

[3] *Email: b.hanna@uq.edu.au Address: School of Languages and Cultures, The University of Queensland, St Lucia QLD 4072, Australia. ORCID ID: https://orchid.org/0000-0002-8450-4243*

[4] *Email: jason.lodge@uq.edu.au Address: School of Education, The University of Queensland, St Lucia QLD 4072, Australia. ORCID ID: https://orchid.org/0000-0001-6330-6160*

[5] *Email: solmaz.abdi@uq.edu.au Address: Institute for Teaching and Learning Innovation, The University of Queensland, St Lucia QLD 4072, Australia. ORCID ID: https://orchid.org/0000-0002-9967-9206*

1. Introduction

As technology increasingly impacts and is used in higher education, relationships between students, staff, and knowledge have changed as the loci of information and expertise diffuse beyond the boundaries of physical campuses. Information is now available anytime, anywhere, with network-enabled devices. This transition to digitally mediated higher education has

significant implications for how students engage with ideas and progress through their studies. In particular, students are increasingly required to engage in self-directed learning, managing many elements of their own study. Active learning, blended learning, and "flipped classrooms" all emphasize student interaction and study outside the classroom, which, in turn, often require students to work on various learning activities, either independently or with peers, without close monitoring by teachers. This means students need to be able to make good decisions about the work they produce and consume.

The implications of the increased impact of technology are perhaps most evident in the enormous body of literature on the importance of self-regulated learning (Panadero, 2017). In this context, there is increasing emphasis on how students can be empowered to reliably and independently judge the quality of their own work and that of others (Cowan, 2010; Sadler, 1989). While the term *evaluative judgment* has recently emerged in higher education to refer to the capacity to make such assessments (Tai, Ajjawi, Boud, Dawson, & Panadero, 2018), the concept itself has been around for at least four decades. Various scholars have urged educators to cultivate students' capacity to discern quality in accordance with an appropriate standard (Hastie & Dawes, 2010; Sadler, 1989). However, although recommendations as to the development of evaluative judgment abound, examples of intervention studies or other kinds of empirical studies are lacking.

This paper aims to fill this research gap by offering a conceptual framework and a referenced implementation for data-driven approaches to investigating the enhancement of evaluative judgment. The framework consists of seven main considerations that should guide development of evaluative judgment tools for learners, educators, and researchers. The referenced implementation demonstrates how the framework can be applied in the ongoing development of an actual educational tool, RiPPLE, that partners with students to create large repositories of learning resources and study material. We present two case studies using RiPPLE to show how a tool that is coherent with the proposed framework can enable the needed empirical research on evaluative judgment. The first is an observational study that investigates how students' evaluations of learning resources compare with those of domain experts, that is, the extent to which their understanding of quality aligns with standards expected by experts. The second presents a randomized controlled experiment that investigates the impact of rubrics on students' evaluations, that is, whether and how the provision of standards in a rubric impacts their rating behaviour, which indicates their evaluative judgment. Our broader aim is to provide a foundation on which to build robust further investigations into optimal support of the development of evaluative judgment in digital environments.

The remainder of this paper is organized as follows. Section 2 provides a brief background to the emergence of the notion of evaluative judgment as relevant to higher education, sketches existing recommendations of approaches to the development of evaluative judgment in students, and highlights the role of technology as a possible facilitator both of students' development of the skill and of the empirical research that is currently lacking. Section 3 presents our conceptual model and RiPPLE as a referenced implementation. Section 4 then demonstrates the research value of this tool through the two aforementioned case studies, and Section 5 concludes the paper.

2. Background

2.1 The Evolution of Evaluative Judgment

In his seminal work on formative assessment and curriculum design, Sadler (1989) specifically emphasized the desirability of students' involvement in assessment as a means to develop their "evaluative knowledge" or "evaluative expertise." He further argued that students' evaluative capabilities should be enhanced so that they can self-monitor their learning processes. In order to do this, students need to understand and appreciate quality.

Subsequent studies introduced similar terms, such as *sustainable assessment* and *informed judgment*, arguing for students to be actively involved in the assessment process (Boud, 2000; Boud & Soler, 2016; Cowan, 2010). For example, the work by Boud (2000) on sustainable assessment promotes the engagement of students in activities designed to improve their capacity to assess the quality of work, equipping them for lifelong learning. Supporting and developing Sadler's concept of evaluative expertise, researchers have called for students to be agentic in the assessment process (Nicol & Macfarlane-Dick, 2006; Nicol, 2010). For instance, Nicol (2010) criticizes monologic delivery of feedback and advocates for feedback that progressively enables students to evaluate their learning/production with little to no input from instructors. In later work, Sadler (2010) emphasizes the need for students to grasp and apply concepts such as the assessment genre, quality, and criteria in order to make complex appraisals of their output. He posits that this allows students to objectively compare the quality of their work to a higher standard and also to develop strategies they can rely on to make improvements (Sadler, 1989). Such student involvement in the understanding and application of standards is in line with the current definition of evaluative judgment as "the capability to make decisions about the quality of work of oneself and others" (Tai et al., 2018, p. 471).

ISSN 1929-7750 (online).

2.2 Developing Evaluative Judgment

An ongoing challenge in higher education is therefore how to effectively assist students in developing elements of evaluative judgment, particularly in the ubiquitous digital environments in which they operate (Lodge, Kennedy, & Hattie, 2018; Khosravi, Gyamfi, Hanna, & Lodge, 2020) and when the curriculum is already stretched to provide both discipline-specific and transferable knowledge and skills.

The following principles have been proposed as paramount to the exercise and development of evaluative judgment: there must be an understanding and application of quality; these should be in reference to an explicit or implicit standard; students should have both opportunities to make judgments that are meaningful in their context and the desire to do so (Tai et al., 2018). With reference to these principles, Sadler (2009) contends that "bringing students into a knowledge of standards requires considerably more than sending them one-way messages through rubrics, written feedback or other forms of telling" (p. 822). In other words, students should be assigned an active role in which they construct knowledge while applying standards. Simply providing students with assessment standards and criteria is not enough for learners and staff to develop a common understanding of them or for students to develop the capacity to apply them accurately. Therefore, in order to develop students' competence at making judgments, it is important that they have opportunities to identify and apply standards (Boud & Soler, 2016).

How then can evaluative judgment best be nurtured? Some recent studies have focused on existing practices that may hold the potential for promoting students' evaluative ability (Ajjawi, Tai, Dawson, & Boud, 2018; Tai, Canny, Haines, & Molloy, 2016; Panadero, Broadbent, Boud, & Lodge, 2019). These include the use of self- and peer assessment (see discussion below), feedback and reviews, rubrics, and exemplars (Tai et al., 2018; Ajjawi et al., 2018). The contention is that these assessment-related activities can enhance the development of students' evaluative judgment by presenting standards of quality and requiring that students apply these standards to their work (Boud & Soler, 2016; Sadler, 2010). But how effective are any of the suggested approaches at fostering evaluative judgment? Much of the existing research on measuring the growth of evaluative judgment has relied on self-report surveys from students (Tai et al., 2016; McConlogue, 2012, e.g.). However, using subjective measures to acquire information about psychological constructs raises concerns relating to cognitive biases and internal validity, since the accuracy of responses cannot easily be verified (Jahedi & Méndez, 2014). The introduction of initiatives to promote the development of evaluative judgment should be grounded in and subject to rigorous empirical research. Furthermore, data on the impact of these initiatives must be collected, and evaluative judgment itself needs to be promoted in time-efficient ways that are compatible with students' other activities. In large part, the unfulfilled need for empirical testing is due to the lack of tools with the assessment approaches required to facilitate large-scale evaluations.

2.3 Educational Tools and Educational Research

Technology provides promising avenues to support students' exercise and development of evaluative judgment using the proposed strategies. A comprehensive review of educational technologies supporting peer assessment by Luxton-Reilly (2009) shows diverse ways in which these tools have incorporated such strategies. While most work concerns peer review (Pirttinen, Kangas, Nygren, Leinonen, & Hellas, 2018; Sung, Chang, Chiou, & Hou, 2005), some tools support the use of guiding rubrics for self- and peer assessment (De Raadt, Toleman, & Watson, 2005; Luxton-Reilly, Plimmer, & Sheehan, 2010; Sondergaard, 2009), and some tools have relied on instructor- and peer-generated exemplars (Bhalerao & Ward, 2001; Cho & Schunn, 2007; Hamer, Kell, & Spence, 2007) as guides for students' production and judgment.

A key benefit of using educational technologies is their capacity to collect rich digital traces on students' behaviour and interactions with the system. The learning analytics and educational data mining communities have explored methods of utilizing such data for research in the context of understanding and enhancing learning, with both fields contributing to the emergence of data-intensive approaches to education (Papamitsiou & Economides, 2014; Siemens & Baker, 2012). However, in most cases educational technologies are built without the aim of supporting research: the platforms themselves do not enable data harvesting or the implementation of observational or controlled experiments. There are, however, successful exceptions: two well-known products are PeerWise (Denny, Hamer, Luxton-Reilly, & Purchase, 2008) and ASSISTments (Heffernan et al., 2016). As a teaching and learning tool, the first allows students to create multiple-choice questions (MCQs); the second is an adaptive assessment tool mostly focused on secondary school mathematics. In terms of facilitating educational research, PeerWise has supported over 80 publications, mainly examining the impact of gamification and the ability of students to develop high-quality learning resources (Denny, 2019). ASSISTments has enabled 27 publications, primarily treating adaptive learning and the personalization of feedback (Heffernan, 2019). This success in enabling research can be attributed to slightly different approaches, both of which make it possible for investigators to see student progress and performance, rather than simply a final outcome. PeerWise allows instructors using the platforms to access data from their courses, rather than the developers retaining exclusive rights to it; the ASSISTments Ecosystem supports purposeful experimental design using randomized controlled trials (RCTs) at low cost (Heffernan et al., 2016). The referenced implementation of the proposed framework, inspired by the success

of these tools, supports empirical educational research by enabling instructors to conduct sound, large-scale, randomized, quasi-experimental, and observational experiments.

In addition, there is a growing consensus that optimal methods of discovering, monitoring, understanding, improving, and evaluating educational processes require the availability of high-quality learning data beyond clickstreams that not only trace student behaviour but also regularly capture their knowledge gains and present learners with the capacity to express their opinions and provide feedback. While there exist exemplar studies and technologies that have organized sequences of learning activities embedded with student responses, providing a comprehensive representation of a learner's learning process (Winne et al., 2019; Holstein, McLaren, & Aleven, 2019; Khosravi & Cooper, 2018), a common understanding and standards for building such systems are still largely undeveloped.

3. Conceptual Model and a Referenced Implementation

This section presents a conceptual framework for developing educational tools that foster and support empirical research on evaluative judgment, together with a referenced implementation using a tool called RiPPLE (Khosravi, Kitto, & Joseph, 2019; Khosravi, Sadiq, & Gasevic, 2020). Figure 1 presents the conceptual framework, consisting of seven main considerations that should guide tool development: (1) Who are the main stakeholders in the tool? (2) What is the main purpose of the tool? (3) Which evaluative judgment (EJ in the figure) strategies can be supported by the tool? (4) What student measurements and metrics are captured by the tool? (5) Which experimental designs are supported by the tool? (6) What ethical design guidelines and principles are relevant to the development of the tool? (7) How are users' interactions with the tool stored?

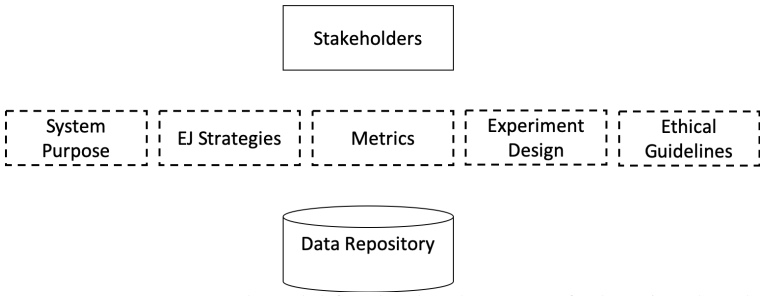


Figure 1. A conceptual model for the development of educational tools to foster and support empirical research on evaluative judgment.

Figure 2 presents an overview of the referenced implementation based on the framework given in Figure 1. In the remainder of this section, we present each component of the framework, discussing them in the context of that implementation.

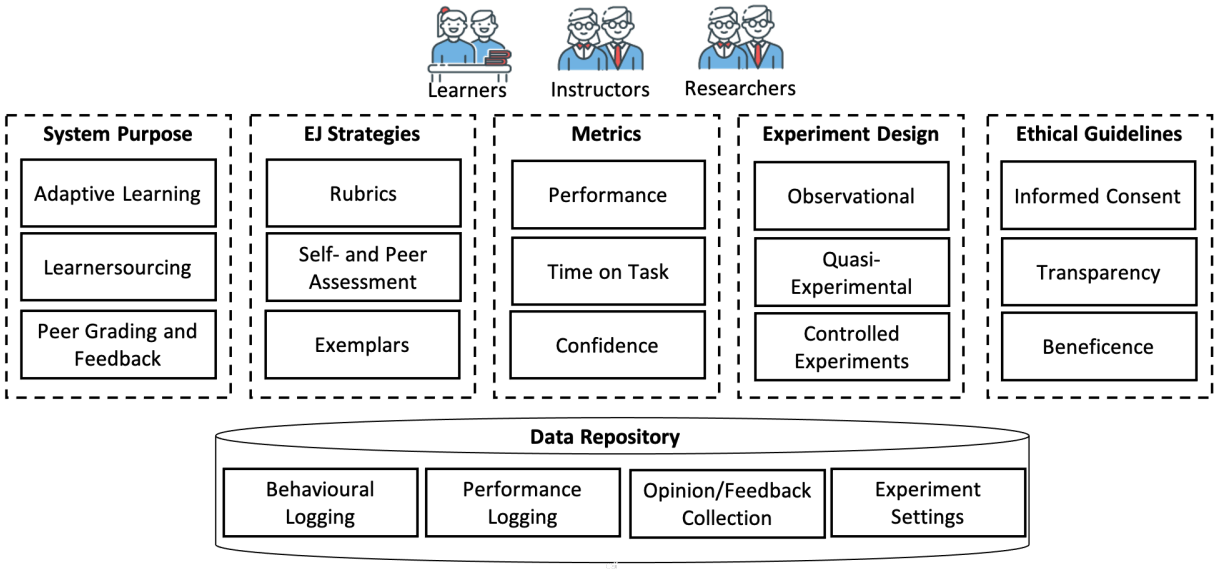


Figure 2. An overview of RiPPLE's implementation based on the proposed conceptual model.

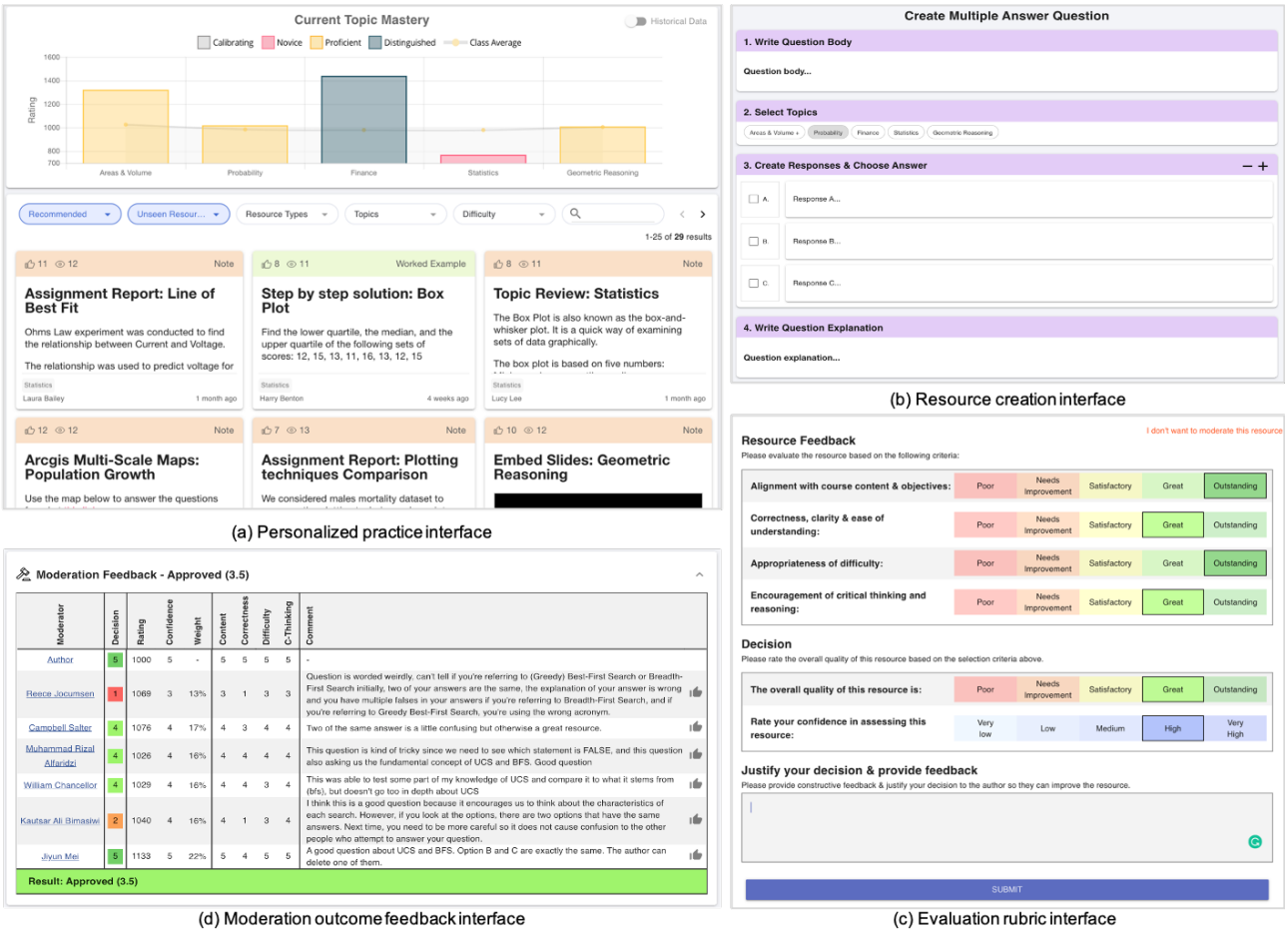
3.1 Stakeholders

A tool should simultaneously serve the needs of a defined set of stakeholders (in the case of evaluative judgment, these could be learners, teachers, researchers, or authorities who set professional standards), represented by the top box in the conceptual model in Figure 1. In the case of RiPPLE, three stakeholder groups are relevant to the design: learners, instructors, and educational researchers. To engage learners, RiPPLE incorporates open learner models to promote self-regulation (Abdi, Khosravi, Sadiq, & Gasevic, 2020). Additionally, RiPPLE has been designed explicitly to put the instructors in charge and empower them with rich analytics (Khosravi, Sadiq, & Gasevic, 2020) and to enable educational research, as further discussed in Section 4.

3.2 System Purpose

Such a tool should be more than a data collection instrument: its purpose(s) must be aligned with educational theory and practice to facilitate and support learning while giving its users the opportunity to make quality judgments of their own work and that of others. In the case of RiPPLE, it aims to support adaptive learning, learnersourcing, and peer grading and feedback, as discussed below.

Adaptive learning. At its core, RiPPLE is an adaptive educational system (Park & Lee, 2004) that dynamically adjusts the level or type of instruction based on individual student abilities or preferences to provide a customized learning experience (Khosravi et al., 2019). Figure 3(a) shows the adaptive practice interface in RiPPLE. The upper part contains an interactive visualization widget allowing students to view an abstract representation of their knowledge state based on a set of topics associated with a course offering. Currently, the approximation of users' knowledge states is produced from an Elo-based rating system (Abdi, Khosravi, Sadiq, & Gasevic, 2019), with the results translated into coloured bars. The lower part of the RiPPLE screen displays learning resources selected for a student based on their learning needs using the recommender system outlined by Khosravi, Cooper, and Kitto (2017).



(a) Personalized practice interface

(b) Resource creation interface

(c) Evaluation rubric interface

(d) Moderation outcome feedback interface

Figure 3. Overview of student modelling and recommendation page of RiPPLE.

Learnersourcing. To provide customized learning for students with different knowledge states, adaptive educational

systems require large repositories of learning resources, which are commonly created by domain experts. Such systems are therefore expensive to develop and challenging to scale. Instead of relying on domain experts as developers, RiPPLE uses a learnersourcing approach to engage students in the creation of learning resources. Figure 3(b) presents the RiPPLE interface for creating a multi-answer question. Not only does this reduce the cost of content generation, but it also holds the potential to foster students' higher-order skills, such as evaluative judgment (Doroudi et al., 2018; Khosravi, Demartini, Sadiq, & Gasevic, 2021). However, since students are still developing their expertise, and even though they are encouraged to reflect on the quality of their work, it is likely that some learning resources created will be ineffective, inappropriate, or incorrect (Bates, Galloway, Riise, & Homer, 2014). Hence, a moderation process is needed to vet the quality of each resource. Here again, RiPPLE relies on the wisdom of the crowd and seeks input from students as moderators, requiring them to judge the quality of their peers' work.

Peer grading and feedback. RiPPLE enables students to moderate the peer-created exercises. Moderation involves indicating the perceived quality of a resource using a five-point Likert scale and providing accompanying written feedback to justify the score. Figure 3(c) presents the RiPPLE interface used to evaluate resources. RiPPLE provides two options for determining the outcome of the evaluation process: (1) instructors act as meta-reviewers and make the final call considering students' evaluations or (2) RiPPLE uses crowd consensus algorithms, as discussed in Darvishi, Khosravi, and Sadiq (2020), to automatically make a decision based on the scores attributed. Figure 3(d) presents the RiPPLE interface for communicating the outcome of the evaluation process. Resources that pass moderation are added to the existing pool of study materials; those that fail are returned to their authors for revision.

3.3 Strategies

An evaluative judgment tool should be able to implement various proposed strategies to make these judgments. These might include marking rubrics of varying degrees of complexity, grading schemes, or open-ended questions. RiPPLE supports the use of a range of these strategies, as discussed below.

Self- and peer assessment. The creation and quality control of the repository of learning resources mean that students are continuously involved in processes of self- and peer assessment, which have been suggested as a means to promote evaluative judgment (Lipnevich, McCallen, Miles, & Smith, 2014; Reddy & Andrade, 2010; Tai et al., 2018). These processes can then be supported by further strategic interventions.

Rubrics. RiPPLE supports the use of rubrics so that a set of defined criteria can guide students' decision-making in self- and peer assessment. Figure 5(b) (below) presents a sample rubric that can be used to moderate resources in RiPPLE. Items used are (1) descriptors of quality based on alignment of the resource with course content, correctness of the resource, and clarity of the resource; (2) overall rating; and (3) students' rating of confidence in their judgment. Self-assessment requires students to use the rubric to evaluate the quality of a resource they have created and only to submit for peer moderation those resources that they believe meet the required standard. In peer assessment, students use the rubric to evaluate the quality of the resources created by their peers.

Exemplars. A further suggested strategy for the development of evaluative judgment is the use of exemplars of acceptable standards (Lipnevich et al., 2014; Carless, Chan, To, Lo, & Barrett, 2018). RiPPLE currently provides these through two general guides. One focuses on content creation and is displayed the first time students access the "Create" tab. It references exemplary learning activities and discusses characteristics of an effective learning resource. The second guide supports content moderation and is shown to students on their first use of the "Moderate" tab. It presents the characteristics of an effective moderation submission.

So far, we have canvassed those features of RiPPLE of most practical interest to the stakeholder groups of learners and instructors: the creation and moderation of learning resources and the implementation of various strategies to raise awareness of quality standards and thus to enhance the development of evaluative judgment. However, as stated above, empirical research on the effect size of each of these interventions is needed. The use of technology-enhanced teaching approaches allows for designs that facilitate such research.

3.4 Metrics

An evaluative judgment tool should be able to capture measurements and metrics related to student progress that can be used to approximate the effectiveness of the proposed strategies. To adapt to the needs of students and provide sufficient information for conducting empirical educational research, RiPPLE enables data collection on various metrics. Three examples closely related to evaluative judgment are discussed below.

Performance. RiPPLE captures students' performance as authors and moderators of resources. This information can be used to determine the effectiveness of different evaluative judgment strategies.

Time on task. An important metric to help researchers understand student learning behaviour and strategies is time on task (Kovanovic, Gašević, Dawson, Joksimovic, & Baker, 2015). RiPPLE uses Google Analytics to collect clickstream data, which is mined to infer the time spent by a student on a particular task. This allows researchers to explore relationships between time spent by a student on a moderation task and the quality of that moderation.

Confidence. The relationship between self-confidence and learning has been widely studied in higher education and in relation to evaluative judgment (Andrade & Brown, 2016; Joughin, Boud, & Dawson, 2019). The rubric used to moderate resources also collects data on students' confidence in their judgment.

3.5 Experimental Design

Tools designed to support educational research should accommodate a range of experimental designs, including A/B tests that allow researchers to conduct controlled experiments.

RiPPLE allows researchers (who may also be instructors) to conduct sound, large-scale studies using observational studies, randomized controlled experiments, and quasi-experimental approaches, as discussed below.

Observational studies. Observational studies are enabled by the provision of access to detailed analytics about student engagement (e.g., access to the platform, moderations performed, ratings provided, comments written) and performance (e.g., resources created, questions answered), through a set of interactive visualizations. For examples of observational studies conducted through RiPPLE, see Khosravi and colleagues (2020) and case study 1 presented in Section 4.1.

RCTs. While the opportunities and challenges of using RCTs in education have been a subject of debate (Sullivan, 2011; Styles & Torgerson, 2018), they remain a gold standard test for establishing causality in some fields of educational research. Although quite expensive and time-consuming to run in physical teaching and learning spaces, in the digital world, RCTs can be cheap and fast. RiPPLE allows such experiments. For an example of RiPPLE supporting educational research using an RCT, see Abdi and colleagues (2020) and case study 2 discussed in Section 4.2.

Quasi-experiments. To help mitigate the ethical challenges of using RCTs in education, RiPPLE also supports quasi-experimental studies where students self-select whether to engage with an intervention. Quasi-experiments are, however, often subject to threats to internal validity: self-selected engagement with an intervention might be influenced by specific traits or needs, meaning that students in the control group are not comparable to those in the experimental group at baseline. However, propensity score matching (PSM) (Rosenbaum & Rubin, 1983) may be applied to reduce baseline differences between the two groups. This method matches each student in the experiment group with a similar student from the control group, with judgments of similarity based on a set of covariates, including features of student performance (e.g., GPA), demographic (e.g., age), and behavioural engagement (e.g., learning management system logins). For an example of a quasi-experimental study using RiPPLE, refer to Khosravi and colleagues (2019).

3.6 Ethical Guidelines

The fifth central component of our conceptual model relates to ethical guidelines. Educational tools should permit the ethical collection of data, with participants providing informed consent within the system itself. The ethical considerations related to the use of student and educational data have been well studied in the field of learning analytics (Drachsler et al., 2015; Ferguson et al., 2014; Pardo & Siemens, 2014). A recent discussion paper (Corrin et al., 2019) highlights the importance of careful handling of student data, providing insightful guidelines, protocols, and principles. Again we use the example of RiPPLE to show how an educational tool may be built to ensure effective compliance with these guidelines.

Consent. On first use of the platform, users are presented with a consent form seeking permission to use their data to improve the academic developers' understanding of the learning process. RiPPLE allows users to change their response at any time. Regardless of their decision, all users can access the platform; however, only data collected from those who have provided and never withdrawn consent are used for research purposes.

Transparency. The platform provides a generic consent form that, in the interests of transparency, must be modified to reflect the specificities of the purpose, scope, and details of planned research.

Beneficence. The terms of use for RiPPLE warn against conducting research involving interventions that may harm students' performance or learning experiences, or simply waste their time.

3.7 Data Repository

A key feature of educational technologies is their capacity to store rich data about students' behaviour and interactions with the system.

In the case of RiPPLE, the data repository stores information about users' performance, behaviour, and opinions as well information about the experiments being conducted on the platform, as discussed below.

Performance logging. Data on students' performance is based on their responses to assessment questions. This information is used to approximate and visualize the mastery level of students across different topics, as specified by Abdi and colleagues (2019) and Abdi, Khosravi, and Sadiq (2021) and shown in the top section of Figure 3(a).

Behaviour logging. RiPPLE stores information about students' behaviour based on their clickstream data, which can be used to identify and understand their study strategies based on algorithms specified in the literature (Jovanović, Gašević, Dawson, Pardo, & Mirriahi, 2017; Matcha, Gašević, Uzir, Jovanović, & Pardo, 2019; Ahmad Uzir, Gašević, Matcha, Jovanović, & Pardo, 2020).

Opinion/feedback collection. The data repository also holds information collected through the range of features that enables students to express their opinion and provide feedback. For example, students subjectively rate and comment on the resources with which they interact. An embedded survey invites students to provide feedback about the platform itself.

Experiment setting. RiPPLE also logs information about different controlled experiments, including information relating to offerings in the experiments, methods of assigning users to groups, the features related to the intervention, and the name of the codebase that is used for each group. The framework for storing experiments is inspired by methods employed by Google Analytics.

4. Case Studies

We present two case studies using RiPPLE to demonstrate how an educational technology designed according to the principles outlined above may enable us to empirically approach the study of how evaluative judgment develops. Each study has its distinct research questions; however, the findings in relation to these questions are less relevant to the purpose of this paper than illustrating how the tool responds to the questions.

4.1 Case Study 1: An Observational Study to Investigate Students' Evaluative Judgment

This study is an example of how RiPPLE may be used to conduct observational empirical research. The overarching aim of this study is to explore the plausibility of determining the quality of learning resources based on the evaluations provided by students, guided by the following research question: How do students' evaluations of learning resources compare with those of domain experts? Here we provide a brief overview of the study. A fuller account is available in Abdi, Khosravi, Sadiq, and Demartini (2021).

4.1.1 Experimental Setting

This study uses data obtained from piloting RiPPLE in an on-campus course on relational databases at the University of Queensland. During the 13 weeks of the semester the course was running, the 521 students enrolled engaged in four rounds of creating, moderating, and using learning resources (MCQs) at three-week intervals. This work plan was implemented to ensure consistent student engagement with RiPPLE throughout the semester. Students' involvement with these activities was rewarded with marks toward their final result. At the end of week 13, the 521 students had made 87,437 attempts and provided 31,143 ratings on the 2,355 student-authored learning resources that were available in the platform's repository for this course. Students were asked to evaluate the effectiveness of a resource using a five-point scale.

To answer the research question and manage the size of the data set, we selected a subset that incorporated sufficient information on active students with different levels of performance in the course as well as MCQs of varying degrees of quality. To do so, we took the following steps:

1. Students with fewer than 25 responses on MCQs were considered inactive, and were excluded.

2. In accordance with the item analysis method in differentiating students (Matlock-Hetzel, 1997), the remaining 384 students were divided into three groups based on their final score in the course: the highest-scoring 27% of students as high-performing (103 students with mean course grade 90.5 ± 4.4), the lowest-scoring 27% of students as low-performing (103 students with mean course grade 54.8 ± 8.5), and the rest as average-performing (178 students with mean course grade 75.4 ± 5).

3. To select MCQs, first, out of the 2,355 MCQs available in the RiPPLE repository, questions that had received fewer than five ratings from each of the three groups of students were excluded, leaving 1,632 to be used in the analysis.

4. Based on their average rating, the remaining 1,632 questions were sorted in ascending order and then divided into three bins, with each group receiving one third of the questions (544 questions in each bin): low-quality questions with mean rating of 2.85 ± 0.44, average-quality questions with mean rating of 3.52 ± 0.13, and high-quality questions with mean rating of 4.1 ± 0.3.

5. Finally, 14 questions from each of these bins were randomly sampled for inclusion in the study, giving a total of 42 questions. Table 1 summarizes the number of ratings given by each student performance group to each bin of learning resources included in this study.

Table 1. Total number of ratings provided by each student performance group for each bin of questions

	Low-performing	Average-performing	High-performing
High-quality	144	411	322
Average-quality	105	286	237
Low-quality	106	216	244

In addition, six individuals with expertise in the course content were recruited as domain experts to independently review the questions. The intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) was used to evaluate the inter-rater agreement among the domain experts. The result of this analysis suggested an excellent agreement among the domain experts for evaluating the quality of learning resources ($ICC[3,k] = 0.84$). Accordingly, for the rest of this study, the domain experts' mean rating for a question is considered as the ground truth for the quality of that question. Table 2 presents the mean and standard deviations of ratings provided by each student group, the entire class, and the domain experts on each bin of the questions.

Table 2. Average and standard deviations of ratings for the three bins of questions by each group of students and domain experts

	Low-performing	Average-performing	High-performing	Class	Experts
High-quality	4.22 ± 1.04	4.23 ± 1.01	4.41 ± 0.85	4.30 ± 0.95	4.38 ± 0.32
Average-quality	3.92 ± 1.19	3.84 ± 1.20	3.87 ± 1.06	3.86 ± 1.16	3.55 ± 0.65
Low-quality	3.16 ± 1.43	3.16 ± 1.31	3.11 ± 1.34	3.14 ± 1.36	2.41 ± 0.61
All resources	3.82 ± 1.29	3.87 ± 1.22	3.85 ± 1.21	3.77 ± 1.16	3.45 ± 0.97

4.1.2 Results and Findings Related to the Research Question

To investigate how students' evaluations of learning resources compared with those of domain experts, we conducted a regression analysis between the ratings provided by the students and those awarded by the domain experts. For this analysis, the student ratings were considered the independent variable and the domain expert ratings the dependent variable. We report the r-value and p-value of the regressed model, in which the r-value represents the Pearson correlation coefficient and the p-value is the two-sided p-value obtained from a Wald test for a hypothesis test for which the null hypothesis is that the slope of the regressed line is zero. The bigger values of Pearson's r correlation suggest a stronger correlation between two variables. In order to examine the significance of the results, we estimate the 95% confidence interval (CI) for the differences between Pearson correlations using the bootstrapping technique. Figure 4 presents the relationship between student ratings and domain expert ratings based on this analysis. In this figure, each data point corresponds to a unique learning resource, and the regressed line represents the best-fit line computed by the regression analysis. At the class level, the results of the Pearson correlation coefficient demonstrate a strong positive correlation between student ratings and those of the domain experts ($r(40) = 0.78, p < 0.01$). In addition, investigations of the student performance group suggest a very strong positive correlation between the ratings provided by the high-performing students and those from the domain experts ($r(40) = 0.828$, $p < 0.01$). The pattern observed for average-performing students is similar to that of high-performing students, but the corresponding data points for this group are more dispersed from the regressed line ($r(40) = 0.694$, $p < 0.01$). However, the difference between the correlation coefficients of high-performing students and low-performing students was not significant with the 95% CI $[-0.01, 0.29]$. Finally, similarly to the other two groups, for low-performing students there is a strong correlation between student rating and domain expert rating ($r(40) = 0.499$, $p < 0.01$). However, the difference between the correlation coefficients of the high-performing and the low-performing students was significant with the 95% CI $[0.17, 0.51]$. In summary, the findings of this study suggest that while there is a strong positive correlation between the ratings provided by the students and the domain experts' ratings, there are evident differences in the ability of students in different performance groups to make judgments.

4.2 Case Study 2: A Controlled Study to Investigate the Impact of Rubrics on Evaluative Judgment

This example demonstrates how RiPPLE may support empirical research on evaluative judgment using a randomized controlled experiment. While the efficacy of rubrics as instructional and assessment tools has been extensively studied in the higher education literature, little empirical research has been carried out on their impact on the development of evaluative judgment (Boud, Lawson, & Thompson, 2013; Bouwer, Lesterhuis, Bonne, & De Maeyer, 2018). This study, therefore, aims to investigate

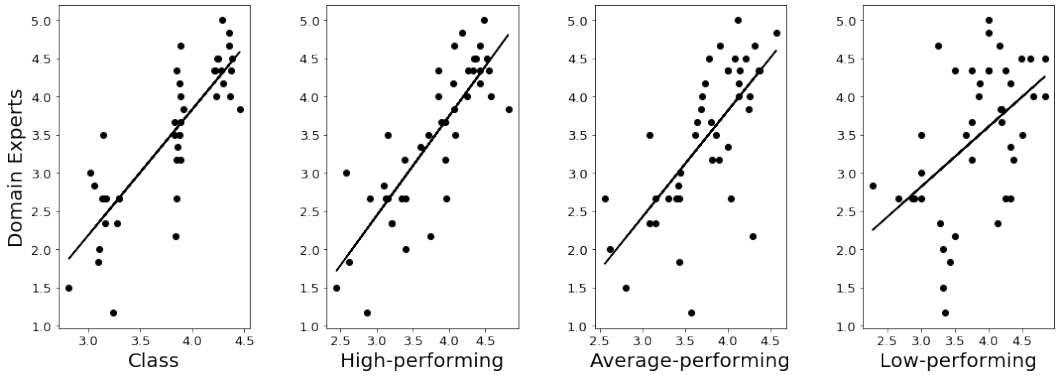
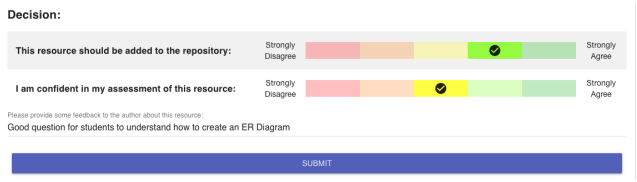


Figure 4. Comparison of the ratings provided by the students and domain experts using the regression analysis.

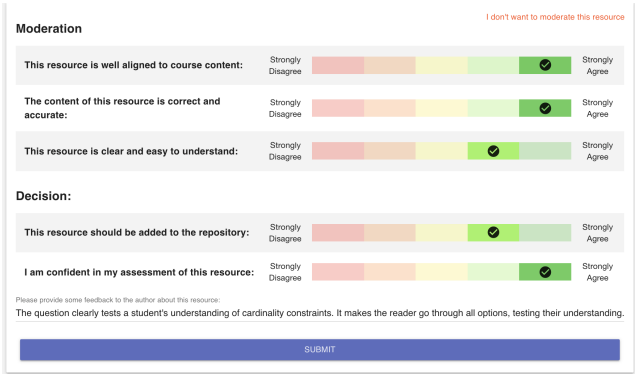
the effects of rubrics on students' moderation of the quality of learning resources deposited to RiPPLE. In particular, we examine the impact of a rubric on ratings, level of agreement on ratings, and confidence level in the assessment of the resources. Here we provide a brief overview of the study. A more in-depth analysis is available in Gyamfi, Hanna, and Khosravi (2021).

4.2.1 Experimental Setting

The study was conducted in an on-campus undergraduate information systems course at the University of Queensland. It used a between-subjects design where 354 participants using RiPPLE were randomly assigned to a control and an experiment group. The control group simply indicated their level of agreement with the proposition that a resource be added to the learning repository, as well as their level of confidence in the accuracy of their recommendation. In contrast, the experiment group was moderated on the basis of a three-item rubric referencing (1) alignment of the resource with course content, (2) correctness of the resource, and (3) clarity of the resource and their confidence in their ratings. Figure 5 shows these two moderation interfaces. Responses used a five-point Likert-scale, where one represents *strongly disagree* and five represents *strongly agree*. In addition to providing numerical scores, participants could provide feedback on and justify their scores in free-text comments. Since each group moderated resources from a different pool of randomly allocated resources, all moderations of a given resource were made by students from the same group.



(a) Control group resource moderation interface.



(b) Treatment group moderation interface.

Figure 5. The control and treatment group moderation interfaces.

Over the five-week period of the study, the control group ($n = 183$) completed 1,143 moderations and the treatment group ($n = 171$) completed 1,069 moderations. A Mann–Whitney statistical analysis was performed on the collected datasets to investigate the effects of rubrics on the process:

- *Ratings*. For each group, participants' quality ratings were collected and analyzed.
- *Confidence*. Participants' confidence ratings in each group were collected and analyzed.
- *Agreement*. For each resource, the standard deviation of the ratings was computed, with a smaller standard deviation

indicating higher agreement. The average (μ) and standard deviation (σ) of deviations across all of the resources were then calculated. The smaller the overall average, the higher the agreement within each group.

- *Length of comments*. The average length of the comments and the standard deviation for each group were computed.

4.2.2 Results

This section reports the results of the four analyses outlined above. Table 3(a) summarizes the findings related to ratings analysis and confidence analysis, and Table 3(b) summarizes the results of agreement and length of comments analysis, where μ and σ stand for the average and standard deviation, respectively.

Table 3. Analysis of data on ratings, confidence, agreement, and length of comment of participants' quality judgment of learning resources from each group. Control $n = 183$ with 1,143 moderations; treatment $n = 171$ with 1,069 moderations.

		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%	
		Control %	**Treatment %**	
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%	

(a) Analysis of data on rating and confidence

	Control μ± σ	Treatment μ± σ	U and P values
Agreement	0.85 ± 0.36	0.80 ± 0.43	U=18470, p=0.271
Length of comments	18.2 ± 19.4	17.9 ± 22.5	U=591332, p=0.191

(b) Analysis of data on agreement and length of comments

Ratings. The results from the ratings analysis reported in Table 3(a) reveal that both participant groups provided positive ratings of resources; however, the treatment group had higher ratings ($\mu = 4.08, \text{Mdn} = 4, \sigma = 1.06$) than the control group ($\mu = 3.94, \text{Mdn} = 4, \sigma = 1.00$); $U = 55{,}038$, $p < 0.01$. We hypothesize that the treatment group provided higher average ratings because the rubric provided was simpler and more generous than the implicit rubric the control group had in mind when moderating. While the rubric provided to the treatment group only asked moderators to consider alignment, correctness, and clarity, many of the comments from the control group moderators indicate that when they rated a resource as ineffective they considered additional factors, such as simplicity and lack of contribution to higher-order learning.

Confidence. The findings from the confidence analysis in Table 3(a) reveal that both groups of students demonstrated high confidence in their ratings (*agree* to *strongly agree*); however, compared to students in the control group ($\mu = 4.22, \text{Mdn} = 4, \sigma = 0.80$), students in the treatment group ($\mu = 4.42, \text{Mdn} = 5, \sigma = 0.88$) showed a statistically significant higher level of confidence in their ratings; $U = 530{,}214$, $p < 0.01$. It seems that the availability of the rubric's explicit presentation of a standard of quality contributed to this greater confidence.

Agreement. The analysis of the average agreement reported in Table 3(b) reveals that participants in the treatment group ($\mu = 0.80, \text{Mdn} = 0.81, \sigma = 0.43$) achieved a slightly higher agreement on quality judgments than participants in the control group ($\mu = 0.85, \text{Mdn} = 0.82, \sigma = 0.36$). However, the difference between the two groups was not statistically significant; $U = 18{,}470$, $p = 0.271$. It is hypothesized that a more comprehensive rubric that better reflected students' implicit expectations of an effective resource could support even higher inter-moderator agreement.

Length of comments. The analysis of the length of comments, measured in words, in Table 3(b) reveals that comments from the control group ($\mu = 18.2, \text{Mdn} = 12, \sigma = 19.4$) were slightly longer than those of the treatment group ($\mu = 17.9, \text{Mdn} = 11, \sigma = 22.5$); however, the difference between the two results was not statistically significant; $U = 591{,}332.5$, $p = 0.191$. This finding demonstrates students' belief that the rubric did not do away with a need for comments.

Overall, the findings suggest that the use of rubrics can significantly impact students' judgment in rating the quality of learning resources and increase their confidence in their ability to do so. However, the high percentage of ratings on the highest point on the scale (five) for the quality of resources suggests that the rubrics encouraged more leniency in marking. In addition, the use of rubrics may also lead to a higher agreement among students judging the quality of a resource, although this increase was not statistically significant. We suggest that the fact that the rubric did not have significantly increased agreement between

and within the groups could be explained by the existence of a shared intrinsic standard of quality that students already had in mind. Furthermore, the use of rubrics did not affect the length of students' comments.

While these findings are interesting, future research should replicate this study across different disciplines with a larger number of students to evaluate the generalizability of the findings.

5. Conclusion and Future Work

In recent years, evaluative judgment, as a key component of self-regulated learning, has been widely discussed in higher education, as have strategies to foster its development. The challenge for educators and researchers has been to study the impact of these strategies, at scale, and without making further onerous demands on themselves and their students in an already crowded curriculum. Despite extensive theoretical work, robust empirical evaluation of the effectiveness of strategies to nurture evaluative judgment has been lacking due in part to lack of tools and technologies that enable appropriate data harvesting. In response to this limitation, this paper has presented a conceptual model as a way of thinking through what would constitute educational tools that can potentially foster students' evaluative judgment while providing a means of conducting large-scale empirical studies. Seven considerations were elaborated upon: stakeholders (learners, instructors and researchers), system purpose, evaluative judgment strategies, metrics, experimental design, ethical guidelines, and storage of data. RiPPLE, a platform designed to support learnersourcing of learning resources from specific cohorts of students, was presented as an example of a tool consistent with the framework. In terms of learning and teaching, it supports currently recommended strategies for developing evaluative judgment, including self- and peer assessment, rubrics, and exemplars. What differentiates RiPPLE from other educational technologies that aim to foster evaluative judgment is its capacity to support large-scale ethical educational research. The application of our framework, through RiPPLE, is demonstrated by two case studies involving large student cohorts producing and evaluating learning resources, thus iteratively applying their developing evaluative judgment. These case studies show data harvesting and analysis relating to students' evaluation of peer-authored resources. Our first case study compares students' ratings with those of experts, showing greater alignment between the judgments of experts and those of higher-performing students; the second investigates the strategy of guiding evaluations through rubrics, showing how an experimental group and a control group can be compared on a range of metrics. These findings afford us quantitative insights into what might seem an intangible, qualitative skill. The framework and tool are therefore shown to provide valuable potential for future work on the enhancement of this critical skill, from both a teaching and a research perspective.

Declaration of Conflicting Interest

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

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Abdi, S., Khosravi, H., & Sadiq, S. (2021). Modelling learners in adaptive educational systems: A multivariate Glicko-based approach. In *Proceedings of the 11th International Conference on Learning Analytics and Knowledge* (LAK 2021), 12–16 April 2021, online (pp. 497–503). New York: ACM. doi: 10.1145/3448139.3448189

Abdi, S., Khosravi, H., Sadiq, S., & Demartini, G. (2021). Evaluating the quality of learning resources: A learnersourcing approach. *IEEE Transactions on Learning Technologies*, *14*(1), 81–92. doi: 10.1109/TLT.2021.3058644

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate Elo-based learner model for adaptive educational systems. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (EDM 2019), 2–5 July 2019, Montreal, Quebec, Canada (pp. 228–233). Retrieved from `https://files.eric.ed.gov/fulltext/ED599177.pdf`

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Complementing educational recommender systems with open learner models. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (LAK 2020)*, 23–27 March 2020, Cyberspace.* New York: ACM. doi: 10.1145/3375462.3375520

Ahmad Uzir, N., Gašević, D., Matcha, W., Jovanović, J., & Pardo, A. (2020). Analytics of time management strategies in a flipped classroom. *Journal of Computer Assisted Learning*, *36*(1), 70–88. doi: 10.1111/jcal.12392

Ajjawi, R., Tai, J., Dawson, P., & Boud, D. (2018). Conceptualising evaluative judgement for sustainable assessment in higher education. In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing Evaluative Judgement in Higher Education* (pp. 23–33). London: Routledge.

Andrade, H. L., & Brown, G. T. (2016). Student self-assessment in the classroom. In G. Brown & L. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 319–334). New York: Routledge.

Bates, S. P., Galloway, R. K., Riise, J., & Homer, D. (2014). Assessing the quality of a student-generated question repository. *Physical Review Special Topics—Physics Education Research*, *10*(2), 020105. doi: 10.1103/PhysRevSTPER.10.020105

Bhalerao, A., & Ward, A. (2001). Towards electronically assisted peer assessment: A case study. *ALT-j*, *9*(1), 26–37. doi: 10.1080/09687760108656773

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., Lawson, R., & Thompson, D. G. (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., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413. doi: 10.1080/02602938.2015.1018133

Bouwer, R., Lesterhuis, M., Bonne, P., & De Maeyer, S. (2018). Applying criteria to examples or learning by comparison: Effects on students' evaluative judgment and performance in writing. *Frontiers in Education*, *3*, 86. doi: 10.3389/feduc.2018.00086

Carless, D., Chan, K. K. H., To, J., Lo, M., & Barrett, E. (2018). Developing students' capacities for evaluative judgement through analysing exemplars. In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing Evaluative Judgement in Higher Education: Assessment for Knowing and Producing Quality Work.* London: Routledge.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409–426. doi: 10.1016/j.compedu.2005.02.004

Corrin, L., Kennedy, G., French, S., Shum, S. B., Kitto, K., Pardo, A., … Colvin, C. (2019). *The Ethics of Learning Analytics in Australian Higher Education.* Melbourne, Australia: University of Melbourne.

Cowan, J. (2010). Developing the ability for making evaluative judgements. *Teaching in Higher Education*, *15*(3), 323–334. doi: 10.1080/13562510903560036

Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learnersourcing to inform design loop adaptivity. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S. M. Dennerlein (Eds.), *Proceedings of the 15th European Conference on Technology Enhanced Learning* (EC-TEL 2020)*, Addressing Global Challenges and Quality Education*, 14–18 September 2020, Heidelberg, Germany (pp. 332–346). Cham: Springer International Publishing. doi: 10.1007/978-3-030-57717-9_24

Denny, P. (2019). *PeerWise Publications.* Retrieved from `https://peerwise.cs.auckland.ac.nz/docs/publications/`

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008). PeerWise: Students sharing their multiple choice questions. In *Proceedings of the Fourth International Workshop on Computing Education Research* (ICER 2008), 6–7 September 2008, Sydney, Australia (pp. 51–58). New York: ACM. doi: 10.1145/1404520.1404526

De Raadt, M., Toleman, M., & Watson, R. (2005). Electronic peer review: A large cohort teaching themselves? In *Proceedings of the 22nd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Balance, Fidelity, Mobility-Maintaining the Momentum?* (ASCILITE 2005), 4–7 December 2005, Brisbane, Australia (Vol. 1, pp. 159–168).

Doroudi, S., Williams, J., Kim, J., Patikorn, T., Ostrow, K., Selent, D., … Rosé, C. (2018). Crowdsourcing and education: Towards a theory and praxis of learnersourcing. In *Proceedings of the 13th International Conference of the Learning Sciences* (ICLS 2018), 23–27 June 2018, London, UK (Vol. 2, pp. 1267–1274). International Society of the Learning Sciences (ISLS).

Drachsler, H., Hoel, T., Scheffel, M., Kismihók, G., Berg, A., Ferguson, R., … Manderveld, J. (2015). Ethical and privacy issues in the application of learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (LAK 2015), 16–20 March 2015, Poughkeepsie, New York (pp. 390–391). New York: ACM. doi: 10.1145/2723576.2723642

Ferguson, R., Clow, D., Macfadyen, L., Essa, A., Dawson, S., & Alexander, S. (2014). Setting learning analytics in context: Overcoming the barriers to large-scale adoption. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (LAK 2014), 24–28 March 2014, Indianapolis, Indiana, USA (pp. 251–253). New York: ACM. doi: 10.1145/2567574.2567592

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

Hamer, J., Kell, C., & Spence, F. (2007). Peer assessment using Aropä. In *Proceedings of the Ninth Australasian Conference on Computing Education* (ACE 2007), 30 January–02 February 2007, Ballarat, Victoria, Australia (Vol. 66, pp. 43–54). Australian Computer Society. Retrieved from https://dl.acm.org/doi/10.5555/1273672.1273678

Hastie, R., & Dawes, R. M. (2010). *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*. Sage. Retrieved from https://us.sagepub.com/en-us/nam/rational-choice-in-an-uncertain-world/book231783

Heffernan, N. (2019). *Assistments: As a researcher's tool.* Retrieved from https://sites.google.com/site/assistmentsstudies/all-studies

Heffernan, N., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., & Williams, J. J. (2016). The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education*, *26*(2), 615–644. doi: 10.1007/s40593-016-0094-z

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics*, *6*(2), 27–52. doi: 10.18608/jla.2019.62.3

Jahedi, S., & Méndez, F. (2014). On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization*, *98*, 97–114. doi: 10.1016/j.jebo.2013.12.016

Joughin, G., Boud, D., & Dawson, P. (2019). Threats to student evaluative judgement and their management. *Higher Education Research & Development*, *38*(3), 537–549. doi: 10.1080/07294360.2018.1544227

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, *33*(4), 74–85. doi: 10.1016/j.iheduc.2017.02.001

Khosravi, H., & Cooper, K. (2018). Topic dependency models: Graph-based visual analytics for communicating assessment data. *Journal of Learning Analytics*, *5*(3), 136–153. doi: 10.18608/jla.2018.53.9

Khosravi, H., Cooper, K., & Kitto, K. (2017). RiPLE: Recommendation in peer-learning environments based on knowledge gaps and interests. *Journal of Educational Data Mining*, *9*(1), 42–67. doi: 10.5281/zenodo.3554627

Khosravi, H., Demartini, G., Sadiq, S., & Gasevic, D. (2021). Charting the design and analytics agenda of learnersourcing systems. In *Proceedings of the 11th International Conference on Learning Analytics and Knowledge* (LAK 2021), 12–16 April 2021, Online—Everywhere! New York: ACM. doi: 10.1145/3448139.3448143

Khosravi, H., Gyamfi, G., Hanna, B. E., & Lodge, J. (2020). Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK 2020), 23–27 March 2020, Frankfurt, Germany (pp. 83–88). New York: ACM. Retrieved from https://doi.org/10.1145/3375462.3375532 doi: 10.1145/3375462.3375532

Khosravi, H., Kitto, K., & Joseph, W. (2019). RiPPLE: A crowdsourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, *6*(3), 91–105. doi: 10.18608/jla.2019.63.12

Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Development and adoption of an adaptive learning system: Reflections and lessons learned. In *Proceedings of the 2020 ACM SIGCSE Technical Symposium on Computer Science Education,* 2020, Online. New York: ACM. Retrieved from https://sigcse2020.sigcse.org/online/papers.html

Kovanovic, V., Gašević, D., Dawson, S., Joksimovic, S., & Baker, R. (2015). Does time-on-task estimation matter? Implications on validity of learning analytics findings. *Journal of Learning Analytics*, *2*(3), 81–110. doi: 10.18608/jla.2015.23.6

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science*, *42*(4), 539–559. doi: 10.1007/s11251-013-9299-9

Lodge, J. M., Kennedy, G., & Hattie, J. (2018). Understanding, assessing and enhancing student evaluative judgement in digital environments. In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing Evaluative Judgement in Higher Education* (pp. 86–94). London: Routledge. doi: 10.4324/9781315109251

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education*, *19*(4), 209-232. Retrieved from https://doi.org/10.1080/08993400903384844 doi: 10.1080/08993400903384844

Luxton-Reilly, A., Plimmer, B., & Sheehan, R. (2010). Studysieve: A tool that supports constructive evaluation for free-response questions. In *Proceedings of the 11th International Conference of the NZ Chapter of the ACM Special Interest Group on Human-Computer Interaction* (CHINZ 2010), 8 July 2010, Auckland, New Zealand (pp. 65–68). New York: ACM. doi: 10.1145/1832838.1832849

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 461–470). doi: 10.1145/3303772.3303787

Matlock-Hetzel, S. (1997). *Basic concepts in item and test analysis.* Paper presented at the Annual Meeting of the Southwest Educational Research Association, 23–25 January 1997, Austin, Texas. Retrieved from https://files.eric.ed.gov/fulltext/ED406441.pdf

McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. *Assessment & Evaluation in Higher Education*, *37*(1), 113–123. doi: 10.1080/02602938.2010.515010

Nicol, D. (2010). *The Foundation for Graduate Attributes: Developing Self-Regulation through Self and Peer-Assessment*. Glasgow, Scotland: The Quality Assurance Agency for Higher Education.

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. doi: 10.1080/03075070600572090

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. doi: 10.3389/fpsyg.2017.00422

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (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. Retrieved from https://doi.org/10.1007/s10212-018-0407-8 doi: 10.1007/s10212-018-0407-8

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64. Retrieved from https://www.jstor.org/stable/jeductechsoci.17.4.49

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450. doi: 10.1111/bjet.12152

Park, O. C., & Lee, J. (2004). Adaptive instructional systems. In D. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology* (2nd ed., pp. 651–684). Mahwah, New Jersey, USA: Lawrence Erlbaum Associates. Retrieved from https://psycnet.apa.org/record/2004-00176-025

Pirttinen, N., Kangas, V., Nygren, H., Leinonen, J., & Hellas, A. (2018). Analysis of students' peer reviews to crowdsourced programming assignments. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research,* 22–25 November 2018, Koli, Finland (pp. 21:1–21:5). doi: 10.1145/3279720.3279741

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. doi: 10.1080/02602930902862859

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi: 10.2307/2335942

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. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*(7), 807–826. doi: 10.1080/03075070802706553

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

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

Siemens, G., & Baker, R. S. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge* (LAK 2012), 29 April–2 May 2012, Vancouver, British Columbia, Canada (pp. 252–254). doi: 10.1145/2330601.2330661

Sondergaard, H. (2009). Learning from and with peers: The different roles of student peer reviewing. *ACM SIGCSE Bulletin*, *41*(3), 31–35. doi: 10.1145/1595496.1562893

Styles, B., & Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research—Methodological debates, questions, challenges. *Educational Research*, *60*(3), 255–264. Retrieved from https://doi.org/10.1080/00131881.2018.1500194 doi: 10.1080/00131881.2018.1500194

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285–289. doi: 10.4300/JGME-D-11-00147.1

Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a web-based self- and peer-assessment system. *Computers & Education*, *45*(2), 187–202. doi: 10.1016/j.compedu.2004.07.002

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*(3), 467–481. doi: 10.1007/s10734-017-0220-3

Tai, J., Canny, B. J., Haines, T. P., & Molloy, E. K. (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

Winne, P. H., Teng, K., Chang, D., Lin, M. P.-C., Marzouk, Z., Nesbit, J. C., … Vytasek, J. (2019). nStudy: Software for learning analytics about processes for self-regulated learning. *Journal of Learning Analytics*, *6*(2), 95–106. doi: 10.18608/jla.2019.62.7