Fostering and Supporting Empirical Research on Evaluative Judgement via a Crowdsourced Adaptive Learning System

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

George Gyamfi
The University of Queensland
Brisbane, QLD, Australia
g.gyamfi@uq.net.au

Barbara Hanna
The University of Queensland
Brisbane, QLD, Australia
b.hanna@uq.edu.au

Jason Lodge
The University of Queensland
Brisbane, QLD, Australia
jason.lodge@uq.edu.au

ABSTRACT

The value of students developing the capacity to make accurate judgements about the quality of their work and that of others has been widely recognised in higher education literature. However, despite this recognition, little attention has been paid to the development of tools and strategies with the potential both to foster evaluative judgement and to support empirical research into its growth. This paper provides a demonstration of how educational technologies may be used to fill this gap. In particular, we introduce the adaptive learning system RiPPLE and describe how it aims to (1) develop evaluative judgement in large-class settings through suggested strategies from the literature such as the use of rubrics, exemplars and peer review and (2) enable large empirical studies at low cost to determine the effect-size of such strategies. A case study demonstrating how RiPPLE has been used to achieve these goals in a specific context is presented.

CCS CONCEPTS

• **Information systems → Personalization**; • **Human-centered computing → Collaborative and social computing systems and tools**; • **Applied computing → Computer-assisted instruction**; **Interactive learning environments**.

KEYWORDS

evaluative judgement, crowd-sourcing, student-authored materials, educational technologies

1 INTRODUCTION

Current teaching and learning practices in higher education emphasise the need to engage students in activities targeting many of the higher-level objectives of the cognitive domain of Bloom's taxonomy [6]. In particular, students' development of evaluative judgement, "the capability to make decisions about the quality of work of self and others" [50] p.471, has been recognised as essential for student learning [3, 9, 35]. Evaluative judgement is a skill that allows students to use feedback effectively, to develop expertise in their field and to extend understanding beyond current work to future endeavours, including lifelong learning [9, 51].

Various strategies for nurturing students' development of evaluative judgement have been discussed in the higher education literature: rubrics, exemplars, and the engagement of students in self- or peer- assessment tasks, have all been hailed as potential effective methods of fostering this capacity [10, 51]. The increasing use of technology in education, and in particular in assessment, provides promising avenues to support students' exercise of evaluative judgement using a range of the approaches noted above. However, in most cases, the impact of the strategies employed cannot be evaluated as the educational technologies used do not enable data harvesting or the implementation of experiments. A common challenge is therefore to both deliver and investigate the impact of such strategies in large cohorts, through means that are reliable, sustainable and ethical [50]. In response, this paper provides an example of an educational technology called RiPPLE that aims to foster evaluative judgement in large-class settings while enabling large empirical studies at low cost to determine the effectiveness of the strategies deployed.

At its core, RiPPLE is an adaptive learning system that makes personalised recommendations of activities to learners, based on their knowledge state, drawing from a pool of crowd-sourced learning resources created by the students and their peers. RiPPLE can potentially foster evaluative judgement in large-class settings by engaging students in assessing the quality and effectiveness of these activities. Strategies from the literature, (e.g. use of invitations to assess, explicit rubrics and exemplars) can support students in evaluating resources. A distinctive feature of RiPPLE is its inbuilt capacity to support educational research, allowing instructors to run observational or controlled experiments.

A case study demonstrating RiPPLE's capacity to foster evaluative judgement and support empirical research is presented. This observational study uses data from an introductory course on Relational Database with 521 students at The University of Queensland that piloted RiPPLE to investigate the following two research questions: (RQ1) How does students' subjective evaluation of learning resources compare with that of domain experts? (RQ2) What is the

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375532

impact of practice over time on students' ability to judge the quality of learning resources? Results suggest that students' subjective ratings of the quality of the learning resources have strong correlations with ratings of domain experts and that students' evaluative judgement improves with practice over time.

2 RELATED WORK

While the term Evaluative Judgement has recently emerged as a topic of interest in higher education, the concept represented is not new: labels such as evaluative knowledge and evaluative expertise [43], informed judgement [7] and judgement of usefulness/ goodness/ trustworthiness [21] have been used to refer to the same construct. Additionally, concepts such as self-regulation, critical thinking and meta-cognition overlap with that of evaluative judgement [35]. Tai et al. [50] hold that two fundamental and complementary principles are pivotal to evaluative judgement: firstly, an understanding of what constitutes quality in a particular field, and secondly, the ability to apply that understanding to an assessment of one's own work and that of others. Similarly, Panadero et al. [35] assert the necessity of an understanding of contexts, quality and standards, assessment criteria and expertise to the formation of quality judgements. The ongoing value for lifelong learning of this capacity to identify, justify and apply a standard of quality or criteria in any situation has been underlined [13, 44, 45].

Previous studies highlight that application of a standard of quality through continuous practice improves students' capacity to determine the quality of their work and that of others [19]. However, students have limited opportunities in this regard. Existing assessment practices position the instructor as the sole authority to evaluate the quality of student productions, with students viewed as ill-equipped to make such judgements effectively [26]. How then can students be engaged and supported in making sound evaluative judgements? Previous studies point to a suite of methods that may enhance students' evaluative judgement abilities, identifying five common techniques: self-assessment, peer feedback/review, feedback, rubrics and exemplars [10, 13, 33, 34, 50].

A comprehensive review by Luxton-Reilly [30] on educational technologies shows diverse ways in which the strategies recommended in the literature have been used to support students' learning through technologically-mediated means. While most tools have been used for peer review [40, 49], some support the use of guiding rubrics for self and peer assessment [14, 31, 46]. In relation to the use of exemplars, some tools have relied on instructor and peer generated samples [5, 11, 20] as standards to guide the quality of students' production and judgement.

While the tools mentioned above may directly or indirectly foster the development of students' evaluative judgement, in most cases the impact of the strategies employed cannot be evaluated without additional data collection: 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 examples of educational technologies which do support research: two well-known products are PeerWise [15] and ASSISTments [23]. As teaching and learning tools, the first allows students to create Multiple Choice Questions (MCQs) and 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 focusing on the impact of gamification and the ability of students to develop high-quality learning resources [39]. ASSISTments has enabled 27 publications, primarily looking at adaptive learning and the personalisation of feedback [22]. This success in supporting research can be attributed to slightly different approaches, both of which allow investigators to look at student progress and performance, not only a final outcome. PeerWise allows instructors using the platform for teaching purposes to access data from those courses, rather than the developers retaining exclusive rights to it; the ASSISTments Ecosystem supports purposeful experimental design using Randomised Control Trials (RCTs) at low cost [23]. The tool used in the current paper, RiPPLE, similarly provides access to rich data from courses and permits observational or randomised controlled experiments. Furthermore, since its use in teaching and learning engages students in the moderation of resources, it allows for research into evaluative judgement, which has not been covered by the work based on PeerWise and ASSISTments.

In summary, the importance of evaluative judgement skills requires the development of teaching interventions grounded in robust empirical research. RiPPLE responds to this imperative by using the strategies mentioned in the literature to foster evaluative judgement while enabling instructors to conduct ethical and sustainable empirical educational research.

3 RIPPLE: A CROWDSOURCED ADAPTIVE LEARNING SYSTEM

This section first provides a brief overview of adaptive learning via RiPPLE (for fuller details, see [27, 28]), before outlining how RiPPLE aims to foster evaluative judgement. Finally, it discusses how RiPPLE supports empirical educational research on a variety of topics, including evaluative judgement.

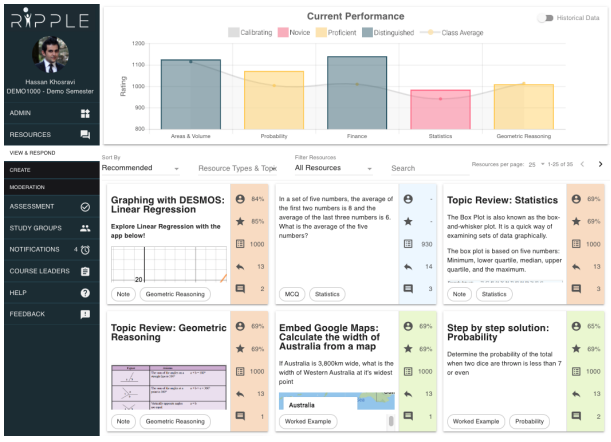

Figure 1: Overview of student modelling and recommendation page of RiPPLE.

3.1 Overview

Adaptive learning systems dynamically adjust the level or type of instruction based on individual student abilities or preferences to

provide a customised learning experience [38]. To support adaptivity, such systems require access to a large repository of learning resources, which are commonly created by domain experts. They are therefore expensive to develop and challenging to scale. Instead of relying on domain experts as developers, RiPPLE uses a crowdsourcing approach to engage students in the creation, moderation and evaluation of learning resources (activities). This not only reduces the cost of content generation, it also carries the potential to foster students' higher-order skills, such as evaluative judgement. To date, over 3000 registered users from 15 courses have used RiPPLE to create roughly 7,000 learning resources, which have been attempted or reviewed over 250,000 times.

Student Modelling and Recommendation. Fig 1 shows one of the main pages in RiPPLE. The upper part contains an interactive visualisation widget allowing students to view an abstract representation of their knowledge state based on a set of topics associated with a course offering. The colour of the bars, determined by the underlying algorithm modelling the student, categorises competence into three levels: for a particular unit of knowledge, red, yellow and blue signify, respectively, inadequate competence, adequate competence with room for improvement, and mastery. Currently, RiPPLE employs the Elo rating system for approximating the knowledge state of users [1]. The lower part of the RiPPLE screen displays learning resources recommended to a student based on his/her learning needs.

Content creation: RiPPLE enables students to create a wide range of learning resources, including MCQs, worked examples, and general notes, incorporating text, tables, images, videos and scientific formulas. Given that students are developing as domain experts, it is likely that some of these learning resources may be ineffective, inappropriate or incorrect [4]. Hence, there is a need for a moderation process to identify the quality of each resource. Here again, RiPPLE relies on the wisdom of the crowd and seeks help from students as moderators.

Content moderation: RiPPLE provides two "formal" moderation options that enable instructors to partner with students to review the quality of the student-created exercises before they are added to a course's repository of learning resources. In both, (1) instructors determine the number of moderations required per resource (e.g., 3 or 5), (2) students review resources and provide a simple judgement (i.e., not effective or effective), alongside a rationale for their decision, (3) instructors determine whether student-creators may appeal the outcome of the moderation. The two moderation options differ as to how the outcome of the process is determined. The two possibilities are (1) majority vote, which is automatically applied, or (2) instructor's final call, based on student moderations. Whichever process is followed, the engagement of students in resource creation and moderation (evaluation) holds potential for fostering evaluative judgement, as discussed below.

3.2 Fostering Evaluative Judgement

To provide effective adaptive learning, RiPPLE relies on students developing and recognising high-quality learning resources, and thus calls on the competencies associated with evaluative judgement: an understanding of quality and the ability to apply it. Herein lies its potential for their development.

Content moderation with the use of rubrics. In its simplest form, moderation only requires students to give a final decision and provide an accompanying rationale (see above). However, students can be provided with a rubric [25, 29, 41, 50] to guide them in their decision-making.

Self-moderation of content with the use of rubrics. This option adds a self-moderation step: once students have created a resource, they are asked to moderate it referring to the rubric used for the formal moderation. Students submit their resource for formal moderation only if their self-moderation determines that it indeed meets the requirements for effectiveness. Again, this aligns with the promotion of skills of evaluative judgement as discussed in the literature [29, 41, 50].

Guidelines using exemplars. Exemplars [10, 29] are currently provided 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 while discussing characteristics of an effective learning resource. The second guide supports content moderation and is shown to students on a first use of the "Moderate" tab. It describes exemplary moderation, presenting characteristics of an effective moderation submission.

Informal evaluation with the use of ratings. Even after a resource has passed formal moderation, students are invited to rate the effectiveness of the resource once they have attempted or reviewed it (see Figure 3). Optionally, the same rubric that was used for formal moderation can be included in this informal evaluation. The case study that is presented in the next section focuses on such informal judgements by students.

The effect size of each of these interventions on fostering evaluative judgement can be investigated using empirical research methods supported by RiPPLE.

3.3 Supporting Ethical Empirical Educational Research

RiPPLE aims to support ethical empirical educational research across large cohorts at low cost.

3.3.1 Ethical Considerations. The ethical considerations bearing on the use of student and educational data have been well studied in the field of learning analytics [17, 18, 36]. A recent discussion paper [12] highlights the importance of careful handling of student data, providing insightful guidelines, protocols and principles. Considerable attention has been given to ensuring the compliance of RiPPLE with these principles. A few examples are given below.

Consent: On their 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 response, all users can access the platform; however, only data collected from learners who have provided and never withdrawn their consent are used for research purposes.

Transparency: The platform provides a generic consent form to researchers and in the interests of transparency, it must be updated to describe any changes to the purpose, scope and details of planned research.

Non-maleficence: The terms of service of using RiPPLE warn researchers against conducting research that leads to interventions

which may harm a student's performance, learning experience, or simply waste their time.

3.3.2 Supporting Empirical Educational Research. Inspired by the success of PeerWise and ASSISTments (see above), RiPPLE aims also to support empirical educational research by enabling instructors to conduct sound, large scale randomised, quasi-experimental and observational experiments. These benefits are discussed below.

Randomised Control Trials (RCTs): While there have been fiery debates about the opportunities and challenges of using RCTs in education [47, 48], 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 enables instructors to conduct such experiments. Currently this must be done through collaboration with the developers, although future versions will allow independent implementation. For an example of RiPPLE supporting educational research using an RCT study please refer to [2].

Quasi-experimental: To help instructors mitigate the ethical challenges of using RCTs in education, RiPPLE also supports quasi-experimental studies where students self-select whether or not to engage with an intervention. Quasi-experiments are 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. Propensity Score Matching (PSM) [42] 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 judgements 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 RiPPLE supporting educational research using an quasi-experiment study please refer to [27].

Observational: RiPPLE also supports observational studies by providing 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 visualisations. Raw data can be read and downloaded as SQL and CSV files. The case study presented below is based on such an observational approach.

4 CASE STUDY

This section provides a case study of how data collected by RiPPLE may be used to conduct empirical research on evaluative judgement. The research questions investigated as part of this demonstration are:

- RQ1. How do students' subjective evaluations of learning resources compare with those of domain experts?
- RQ2. What is the impact of practice over time on students' ability to judge the quality of learning resources?

4.1 Experimental Setting

The data set used for this study was obtained from a pilot of RiPPLE in a 13-week course on Relational Database with 521 participants at The University of Queensland. To ensure consistent engagement and maximise practice in the course, students were involved in 4 rounds of creating and answering MCQs at 3-week intervals. Participation in creating, using and evaluating resources was rewarded with marks towards their final grade. At the end of the course, the participants had made 87,437 response attempts and provided 31,143 ratings on the 2,355 student-authored learning resources (MCQs).

To rate the quality of a resource, students were instructed to consider the following criteria: (1) The question reinforces learning from the content covered in the course; (2) The author has provided a good solution to the question: their explanation must be helpful to someone who answers their question incorrectly; and (3) Other options must seem plausible. The evaluation was performed through the attribution of a global score, represented by a number of stars out of a maximum of five. Figure 2 illustrates the interface used for the evaluation of the resources.

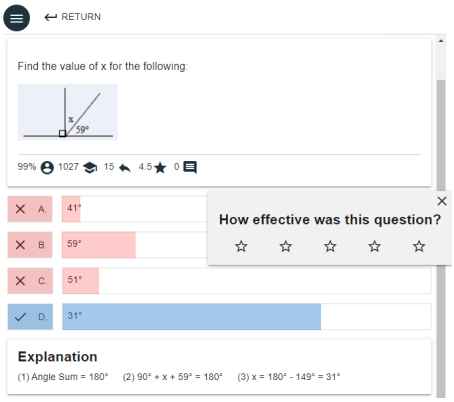

Figure 2: Overview of the interface used for evaluating resources on RiPPLE.

To ensure that the data set contained sufficient information on active high-performing, average-performing and low-performing students as well as high-quality, average-quality and low-quality resources, the following steps were followed: (1) Students who had answered less than 25 questions were considered inactive and were removed from the study; (2) The remaining 384 students were divided into three groups based on their final score in the course. In accordance with Item Analysis in differentiating students [32], the highest-scoring 27% of students were considered as high-performing, the lowest scoring 27% of students as low-performing and the rest as the average-performing; (3) Of the 2,355 MCQs in the platform's repository, the study only considered those which had received more than 10 ratings from each of the three groups (high- average- low- performing); (4) To determine the assessed quality of the remaining 1,632 questions, they were arranged in ascending order by their average ratings and then divided into three groups (low- average- high- quality questions); (5) Fourteen questions were randomly sampled from each of these groups for use in the final study, giving a total of 42 questions. In sum, 319 students, 42 questions and 2070 student ratings were included in the final study. Value aggregations of students' weekly interactions were compiled for analysis.

To provide a point of comparison, six domain experts with expertise in the course content were recruited to review and rate the quality of the 42 questions according to the same criteria the students used. These experts recorded a total of 252 ratings.

4.2 Results and Findings Related to RQ1

To investigate RQ1, we looked for any existing correlation between domain expert ratings and student ratings of the quality of learning resources. A regression analysis treated the domain experts' ratings as the dependent variable and the ratings given by students as the independent variable. In addition to visual explorations, we also report the r and p of the regressed model where r is the Pearson correlation coefficient and p 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.

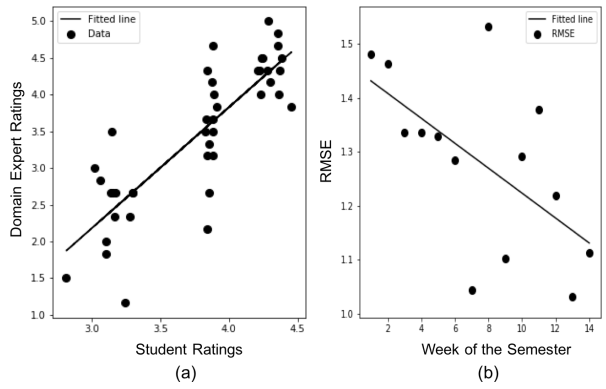

Figure 3: Figure (a) shows the relationship between student ratings and domain expert ratings. Figure (b) shows the average RMSE of student ratings for each week of the semester.

Figure 3(a) presents the results. This analysis reveals a strong and positive correlation (r=0.832, p<0.05) between ratings from the two groups. The results obtained provide evidence that students are effective at assessing the quality of learning resources. This finding aligns with those of [52] which indicated that students have the ability to rate the quality of resources appropriately. The finding also matches those of Denny et. al [16] where ratings of peer-created MCQs were not only found to be valid but also correlated with instructor ratings (r=0.5 and r=0.58). They further asserted that students participated actively, generated and rated the quality of the resources using higher order learning skills. Similarly, a study on peerScholar found a good correlation between student and expert ratings [37]. Furthermore, this result is in line with [24]'s study on students' understanding of multivariate calculus in which high validity and inter-rater reliability were found between evaluative responses of students, experts, novice and marks from other tests. Our study therefore adds to the literature demonstrating students' capacity to make useful evaluative judgements. The question remains however as to whether they can refine their capacity to judge, closing the gap between their ratings and those of experts.

4.3 Results and Findings Related to RQ2

To investigate RQ2, we considered the domain experts' ratings as the gold standard and computed the error of the ratings provided by each group of students on the quality of learning resources. Root Mean Squared Error (RMSE) was used to compute the error of ratings provided by students and was computed as $\sqrt{\frac{\sum_{(i)}(e_i-s_i)^2}{N}}$ where e_i and s_i are the ratings for learning resource i expressed by the domain experts and students, respectively, and N is the number of all resources in the data set for which RMSE is being reported. The RMSE of the ratings provided by students on a weekly basis during the 13 weeks of the semester were used to conduct a regression analysis on the obtained RMSE values throughout the semester.

Figure 3(b) presents the results of this analysis, which demonstrates a statistically significant association between weekly RMSE value at p=0.0226 and a mild inverse correlation (reduction in student error in quality ratings) between week r=-0.602. That is, as weeks increase, the RMSE decreases. Remembering that student participation across semester was incentivised through the assessment protocol, time elapsed equates to practice. This suggests that, over time and with practice, students narrowed the gap between their ratings and those of domain experts, hence providing evidence that students developed an enhanced understanding of the concept of "quality". Our findings align with those of [8] who found that students can become better judges within and across different subjects over time when they engaged in self-assessment using a web-based marking and feedback system, ReView, for two to four semesters. Their quality ratings were in general in line with those of instructors.

In sum, our study affirms the ability of students to develop their evaluative judgement of MCQs served adaptively in a digital environment, simply through repeated exercise. It therefore suggests the value of more systematic interventions to improve students' evaluative judgement and demonstrates the use of RiPPLE as a research tool to verify and measure such interventions.

5 CONCLUSION AND FUTURE WORK

The importance of evaluative judgement in higher education cannot be underestimated. However, there is still work to be done on how to facilitate its development among students in higher education. This paper has provided a demonstration of how educational technologies can contribute, both through facilitating students' opportunities to hone their evaluative judgement and by supporting empirical research in the area. As a learning environment, our example RiPPLE, can deliver strategic interventions recommended in the literature as contributing towards the development of evaluative judgement such as repeated practice of peer moderation; rubrics for assessment; exemplars. An important differentiator of RiPPLE compared to many of the educational tools that foster evaluative judgement is that it is designed with the aim of supporting ethical educational research across large cohorts at a low cost. Promising directions for future research include but are not limited to investigating the effect size and the importance of each of the interventions proposed in the literature on students' capacity for evaluative judgement.

REFERENCES

[1] Solmz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2019. A Multivariate ELO-based Learner Model for Adaptive Educational Systems. In *Proceedings of the Educational Data Mining Conference*. 462–467.

[2] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Complementing Educational Recommender Systemswith Open Learner Models. In *Proceedings of the tenth International Conference on Learning Analytics And Knowledge*. ACM.

[3] Michael Absolum, Lester Flockton, John Hattie, Rosemary Hipkins, and Ian Reid. 2009. Directions for assessment in New Zealand: Developing students' assessment capabilities. *Unpublished paper prepared for the Ministry of Education* (2009).

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

[5] Abhir Bhalerao and Ashley Ward. 2001. Towards electronically assisted peer assessment: a case study. *ALT-J* 9, 1 (2001), 26–37.

[6] Benjamin S Bloom et al. 1956. Taxonomy of educational objectives. Vol. 1: Cognitive domain. *New York: McKay* (1956), 20–24.

[7] David Boud. 2007. Reframing assessment as if learning were important. In *Rethinking assessment in higher education*. Routledge, 24–36.

[8] David Boud, Romy Lawson, and Darrall G Thompson. 2013. Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education* 38, 8 (2013), 941–956.

[9] David Boud and Rebeca Soler. 2016. Sustainable assessment revisited. *Assessment & Evaluation in Higher Education* 41, 3 (2016), 400–413.

[10] David Carless, Kennedy Kam Ho Chan, Jessica To, Margaret Lo, and Elizabeth Barrett. 2018. Developing students' capacities for evaluative judgement through analysing exemplars. *Developing Evaluative Judgement in Higher Education: Assessment for knowing and producing quality work*. (2018).

[11] Kwangsu Cho and Christian D Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education* 48, 3 (2007), 409–426.

[12] Linda Corrin, Gregor Kennedy, Sarah French, Simon Buckingham Shum, Kirsty Kitto, Abelardo Pardo, Deborah West, Negin Mirriahi, and Cassandra Colvin. 2019. The ethics of learning analytics in Australian higher education.

[13] John Cowan. 2010. Developing the ability for making evaluative judgements. *Teaching in Higher Education* 15, 3 (2010), 323–334.

[14] Michael De Raadt, Mark Toleman, and Richard Watson. 2005. Electronic peer review: A large cohort teaching themselves?. In *Proceedings ASCILITE 2005: 22nd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Balance, Fidelity, Mobility-Maintaining the Momentum?*, Vol. 1. Queensland University of Technology, Teaching and Learning Support Services, 159–168.

[15] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research*. ACM, 51–58.

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

[17] Hendrik Drachsler, Tore Hoel, Maren Scheffel, Gábor Kismihók, Alan Berg, Rebecca Ferguson, Weiqin Chen, Adam Cooper, and Jocelyn Manderveld. 2015. Ethical and Privacy Issues in the Application of Learning Analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15)*. ACM, New York, NY, USA, 390–391. https://doi.org/10.1145/2723576.2723642

[18] Rebecca Ferguson, Doug Clow, Leah Macfadyen, Alfred Essa, Shane Dawson, and Shirley Alexander. 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*. ACM, 251–253.

[19] Jon Guest and Robert Riegler. 2017. Learning by doing: Do economics students self-evaluation skills improve? *International Review of Economics Education* 24 (2017), 50–64.

[20] John Hamer, Catherine Kell, and Fiona Spence. 2007. Peer assessment using aropä. In *Proceedings of the Ninth Australasian Conference on Computing Education-Volume 66*. Australian Computer Society, Inc., 43–54.

[21] Reid Hastie and Robyn M Dawes. 2010. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage.

[22] Neil Heffernan. 2019. *ASSISTments: As a researcher's tool.* https://sites.google.com/site/assistmentsstudies/all-studies

[23] Neil T Heffernan, Korinn S Ostrow, Kim Kelly, Douglas Selent, Eric G Van Inwegen, Xiaolu Xiong, and Joseph Jay Williams. 2016. The Future of Adaptive Learning: Does the Crowd Hold the Key? *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 615–644.

[24] Ian Jones and Lara Alcock. 2014. Peer assessment without assessment criteria. *Studies in Higher Education* 39, 10 (2014), 1774–1787.

[25] Sam Kavanagh and Andrew Luxton-Reilly. 2016. Rubrics used in peer assessment. In *Proceedings of the Australasian Computer Science Week Multiconference*. ACM.

[26] Mohammed K Khalil and Ihsan A Elkhider. 2016. Applying learning theories and instructional design models for effective instruction. *Advances in physiology education* 40, 2 (2016), 147–156.

[27] Hassan Khosravi, Kirsty Kitto, and Williams Joseph. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of learning analytics* 6, 3 (2019), 83–97.

[28] Hassan Khosravi, Shazia Sadiq, and Gasevic Dragan. 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*. ACM.

[29] Anastasiya A Lipnevich, Leigh N McCallen, Katharine Pace Miles, and Jeffrey K Smith. 2014. Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science* 42, 4 (2014), 539–559.

[30] Andrew Luxton-Reilly. 2009. A systematic review of tools that support peer assessment. *Computer Science Education* 19, 4 (2009), 209–232.

[31] Andrew Luxton-Reilly, Beryl Plimmer, and Robert Sheehan. 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*. ACM, 65–68.

[32] Susan Matlock-Hetzel. 1997. Basic Concepts in Item and Test Analysis. (1997).

[33] David Nicol. 2014. Guiding principles for peer review: Unlocking learners' evaluative skills. *C. Kreber, C. Anderson, N. Entwistle and J. McArthur, Advances and innovations in university assessment and feedback* (2014), 197–224.

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

[35] Ernesto Panadero, Jaclyn Broadbent, David Boud, and Jason 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 (01 Jul 2019), 535–557. https://doi.org/10.1007/s10212-018-0407-8

[36] Abelardo Pardo and George Siemens. 2014. Ethical and privacy principles for learning analytics. *British Journal of Educational Technology* 45, 3 (2014), 438–450.

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

[38] Ok choon Park and Jung Lee. 2004. *Adaptive instructional systems*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 651–684.

[39] Denny Paul. 2019. *PeerWise Publications*. https://peerwise.cs.auckland.ac.nz/docs/publications/

[40] Nea Pirttinen, Vilma Kangas, Henrik Nygren, Juho Leinonen, and Arto Hellas. 2018. Analysis of students' peer reviews to crowdsourced programming assignments. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. ACM, 21.

[41] Y Malini Reddy and Heidi Andrade. 2010. A review of rubric use in higher education. *Assessment & evaluation in higher education* 35, 4 (2010), 435–448.

[42] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55. https://doi.org/10.2307/2335942

[43] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144.

[44] D Royce Sadler. 2010. Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education* 35, 5 (2010), 535–550.

[45] D Royce Sadler. 2016. Three in-course assessment reforms to improve higher education learning outcomes. *Assessment & Evaluation in Higher Education* 41, 7 (2016), 1081–1099.

[46] Harald Sondergaard. 2009. Learning from and with peers: the different roles of student peer reviewing. *ACM SIGCSE Bulletin* 41, 3 (2009), 31–35.

[47] Ben Styles and Carole Torgerson. 2018. Randomised controlled trials (RCTs) in education research - methodological debates, questions, challenges. *Educational Research* 60, 3 (2018), 255–264. https://doi.org/10.1080/00131881.2018.1500194 arXiv:https://doi.org/10.1080/00131881.2018.1500194

[48] Gail M Sullivan. 2011. Getting off the "gold standard": randomized controlled trials and education research. *Journal of graduate medical education* 3, 3 (2011), 285–289.

[49] Yao-Ting Sung, Kuo-En Chang, Shen-Kuan Chiou, and Huei-Tse Hou. 2005. The design and application of a web-based self-and peer-assessment system. *Computers & Education* 45, 2 (2005), 187–202.

[50] Joanna Tai, Rola Ajjawi, David Boud, Phillip Dawson, and Ernesto Panadero. 2018. Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education* 76, 3 (01 Sep 2018), 467–481. https://doi.org/10.1007/s10734-017-0220-3

[51] Joanna Hong-Meng Tai, Benedict J Canny, Terry P Haines, and Elizabeth 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 (2016), 659–676.

[52] Jacob Whitehill, Cecilia Aguerrebere, and Benjamin Hylak. 2019. Do Learners Know What's Good for Them? Crowdsourcing Subjective Ratings of OERs to Predict Learning Gains. In *Proceedings of the Educational Data Mining Conference*.