Charting the Design and Analytics Agenda of Learnersourcing Systems

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

Gianluca Demartini
The University of Queensland
Brisbane, QLD, Australia
g.demartini@uq.edu.au

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

Dragan Gasevic
Monash University
Clayton, Vic, Australia
Dragan.Gasevic@monash.edu

ABSTRACT

Learnersourcing is emerging as a viable learner-centred and pedagogically justified approach for harnessing the creativity and evaluation power of learners as experts-in-training. Despite the increasing adoption of learnersourcing in higher education, understanding students' behaviour while engaged in learnersourcing and best practices for the design and development of learnersourcing systems are still largely under-researched. This paper offers data-driven reflections and lessons learned from the development and deployment of a learnersourcing adaptive educational system called RiPPLE, which to date, has been used in more than 50-course offerings with over 12,000 students. Our reflections are categorised into examples and best practices on (1) assessing the quality of students' contributions using accurate, explainable and fair approaches to data analysis, (2) incentivising students to develop high-quality contributions and (3) empowering instructors with actionable and explainable insights to guide student learning. We discuss the implications of these findings and how they may contribute to the growing literature on the development of effective learnersourcing systems and more broadly technological educational solutions that support learner-centred learning at scale.

CCS CONCEPTS

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

KEYWORDS

Learnersourcing, crowdsourcing in education, human-centred computing, explainable AI

ACM Reference Format:
Hassan Khosravi, Gianluca Demartini, Shazia Sadiq, and Dragan Gasevic. 2021. Charting the Design and Analytics Agenda of Learnersourcing Systems. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448143

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
LAK21, April 12–16, 2021, Irvine, CA, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448143

1 INTRODUCTION

Contemporary models of learning have emphasised the importance of learner-centred approaches that (1) engage learners in active and higher-order learning activities [4, 40], which enable learners to develop their own vision, reasoning, and judgement to extend understanding, (2) provide rich and timely feedback [22, 68], which enables learners to make sense of information about their performance and use it to enhance their learning strategies, and (3) personalise learning [38, 58], to tailor learning instructions to the individual needs of learners. However, as the number of learners grows, employing the classical approach of having an instructor to facilitate via learner-centred approaches becomes more challenging [48]. One viable and increasingly recognised approach for addressing this challenge is to employ *learnersourcing* [20, 36]. The concept of *learnersourcing refers to a pedagogically supported form of crowdsourcing that mobilises the learner community as experts-in-training to contribute to teaching or learning while being engaged in a meaningful learning experience themselves.*

Previous studies have demonstrated examples of how learnersourcing can be used towards facilitating a learner-centred approach based on the aforementioned three points. Examples of employing learnersourcing to engage learners in higher-order learning tasks includes enabling students to create and evaluate knowledge components [47], multiple-choice questions [17, 34], personalised hints [25], summaries of steps in how-to videos [64], explanations for peer instruction [7], solutions to open-ended questions [63], explanations for programming misconceptions [26], and comparing and contrasting pairs of similar learning artefacts [19]. Learnersourcing is also commonly used to support the delivery of feedback using peer assessment and grading systems (e.g., [52, 67]), which have been demonstrated to help learners develop evaluative judgement, the capacity to make accurate evaluations about the quality of their work and that of others [33, 56]. Furthermore, one of the main applications of learnersourcing has been to support the development and evaluation of content that can be used within adaptive engines to support personification of education [28, 31, 35, 66].

Despite the increasing adoption of learnersourcing systems in higher education, best practices and methods for the design and development of learnersourcing systems are still largely under-researched. This paper aims to contribute to filling this research gap and further promoting the use of learnersourcing within the learning analytics community by sharing data-driven reflections and lessons learned from the design, development, and deployment of a learnersourcing adaptive educational systems called RiPPLE. To date, RiPPLE has been used in more than 50-course offerings with roughly 12,000 students. Our data-driven reflections are targeted

towards highlighted challenges in learnersourcing based on past studies (as discussed in Section 3) and are guided by the following three questions. How can learnersourcing systems: (1) accurately and transparently assess the quality of students' contributions?, (2) be designed to incentivise a large portion of the student population to offer high-quality contributions?, and (3) empower instructors with actionable and explainable insights to provide oversight? We discuss the implications of our findings with a focus on how we may contribute to the growing literature on the development of effective learnersourcing systems and more broadly technological educational solutions that support learner-centred learning at scale. In what follows, Section 2 provides a brief overview of RiPPLE. Section 3 offers our data-driven reflections and lessons learned. Finally, Section 4 provides a brief discussion and concluding remarks.

2 THE RIPPLE SYSTEM

A full description of RiPPLE is provided in [34]. Here, we provide a brief description based on the features of RiPPLE that are relevant to the context of this paper. At its core, RiPPLE is an adaptive educational system (AES) that dynamically adjusts the level or type of instruction based on individual student abilities or preferences [50]. To effectively adapt to the learning needs of individual students, AESs require access to a large repository of learning resources. These resources are commonly created by domain experts [5], which makes AESs expensive to develop and challenging to scale. RiPPLE takes the learnersourcing approach of partnering with students to create a repository of learning resources. Students have the ability to contribute different types of resources including multiple-choice questions, multi-answer questions, matching questions, worked examples, and open-ended notes.

Previous work has shown that the quality of learnesourced content is rather diverse with some developed resources meeting rigorous judgemental criteria while others are ineffective, inappropriate, or incorrect (e.g., [6, 23]). Consequently, to effectively utilise a learnersourced repository of content, there is a need for a selection and moderation process to separate high-quality from low-quality resources. One approach for doing this is to engage instructors as experts in evaluating the quality of the resources; however, the instructor-led quality evaluation is not scalable and can be expensive due to the potentially large size of these repositories. An alternative solution, which is employed by RiPPLE, is to develop a formal evaluation process that again relies on learnersourcing, where students review and evaluate existing resources. RiPPLE follows the process of academic journals and assigns each resource to be evaluated by multiple moderators. However, given the large number of resources generated by students, it is unrealistic to expect instructors to act as a meta reviewer and make a final decision on the quality of resources that have been evaluated by students. Therefore, as discussed in Section 3.1.1, RiPPLE automates the decision making process for inferring the quality of a resource.

Fig 1-a shows the personalised practice interface 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 lower part of the practice interface displays learning resources recommended to a student based on their learning needs using the recommender system outlined in [32]. Fig 1-b illustrates an example of the interface used for creating learning resources. The provided example shows the page used for creating multiple answer questions. Fig 1-c illustrates the interface used by a student or instructor moderator for evaluating a resource. It includes a rubric of four items, which asks the moderator to consider the alignment, correctness, difficulty level, and critical thinking level of a resource. Moderators then provide a final decision and their confidence in their own rating. Moderators are expected to justify their decision and provide feedback to the author before submitting their evaluation. Finally, Fig. 1-d shows an example of how evaluations and the inferred outcome are shared with the author, moderators and instructors. The authors of approved resource are encouraged to update their resources based on the feedback provided. Their resource is added to a repository of resources that are used in the adaptive engine of RiPPLE. The authors of rejected resources can update and resubmit their resource; however, if resubmitted, the resource will be considered as a new submission and will have to go through the moderation process again.

3 DATA-DRIVEN REFLECTIONS AND LESSONS LEARNED

In this section, we synthesise the findings from the literature with reflections and lessons learned from designing, developing and deploying RiPPLE. Our synthesis draws insights from piloting RiPPLE in over 50 courses across a range of disciplines including Medicine, Pharmacy, Psychology, Education, Business, Computer Science and Biosciences with roughly 12,500 students who have authored over 20,000 learning resources and 70,000 formal evaluation and over 1 million interactions on these resources study[1].

3.1 Assessing the Quality of Learnersourced Content with Accurate, Explainable and Fair Approaches

In decision-making tasks, due to the potential that the decision made by an individual might be incorrect, many systems employ a redundancy-based strategy and assign the same tasks to multiple individuals. The problem of optimal integration of the crowdsourced decisions in the absence of a ground truth towards making an accurate final decision has been studied extensively within the crowdsourcing community [72]. Many of the state-of-the-art crowd consensus approaches rely on machine learning algorithms (e.g., ([16]) to simultaneously infer the true outcome and contributors' reliability. Past studies have shown that the use of machine learning algorithms have significantly improved the accuracy of the models compared to averaging aggregation functions [72]; however, these methods often lack understandability and transparency (in terms of how individuals were rated and how a final decision was made). The literature suggests that the use of explainable AI (XIA) [61] is not always wanted or necessary [11]. However, the use of machine learning algorithms with black-box outcomes seems to be particularly inadequate for educational settings where educators strive to provide extensive feedback to enable learners to develop their own

[1]Approval from our Human Research Ethics Committee with id #2018000125 was received for conducting the studies and observations reported.

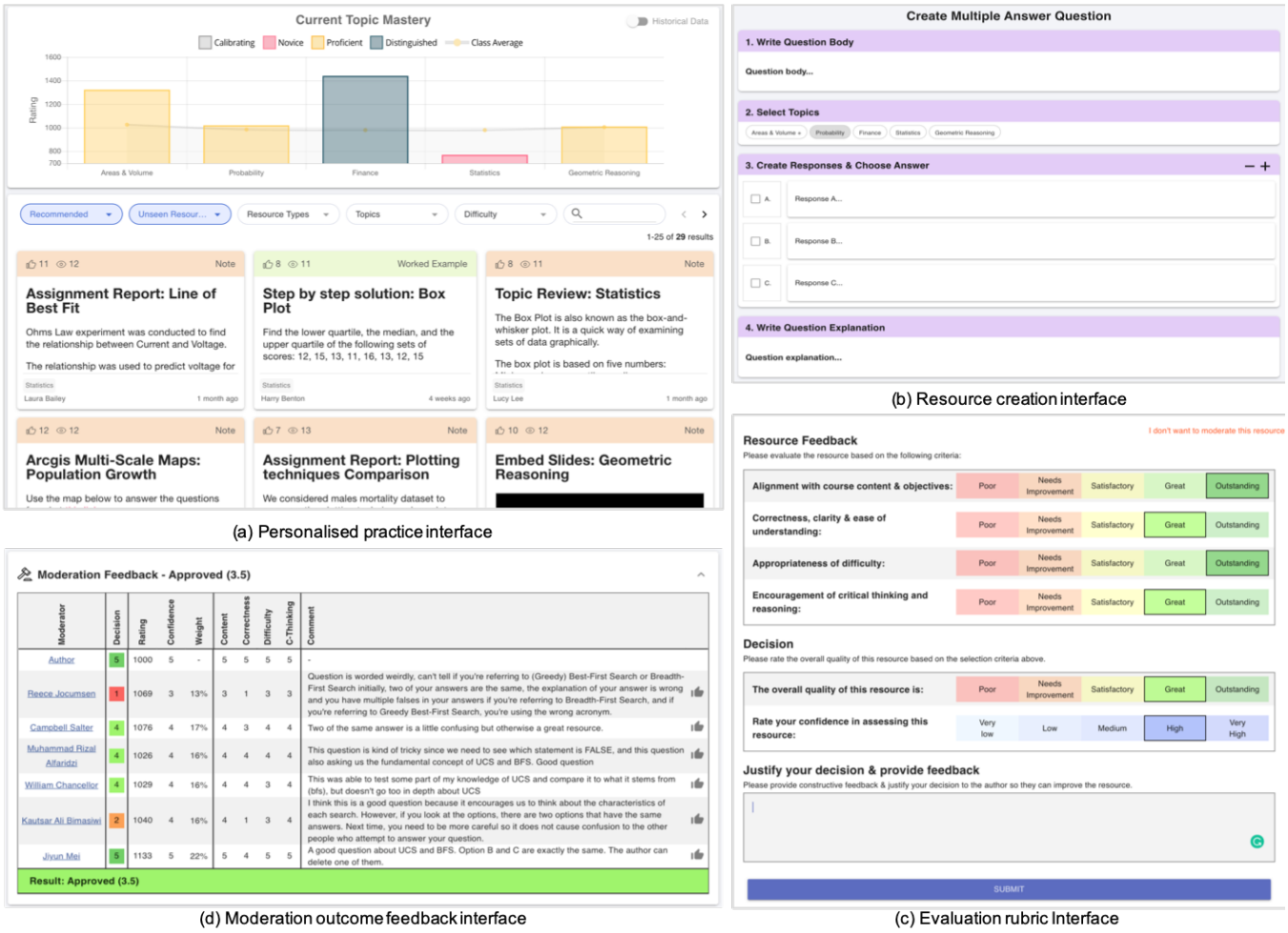


Figure 1: Four of the main interfaces of RiPPLE

vision, reasoning, and appreciation for inquiry and investigation and fairness.

Much of the existing work on the need for open and XIA models in education has been conducted in the field of open learner models [9] where models are often opened through visualisations, as an important means of supporting learning through various systems such as learning analytics dashboards [10, 54], intelligent tutoring systems [53], educational recommender systems [3], and adaptive learning platforms [34] (please see Section 3.3 for further discussion on use of explainable AI in education). In terms of learnersourcing systems, the problem of assessing quality of learnersourced contributions has been referred to or studied in previous work [19, 25, 47, 63]; however the focus has generally been on maximising accuracy rather than explainability. So how can learnersourcing systems employ crowd consensus approaches that are accurate but also explainable? How can they enable students or instructors to raise concerns if they think the outcome is unfair? What criteria should students be asked to consider in their evaluations? The remainder of this section provides data-driven reflections from our attempt on answering these three questions within the RiPPLE system.

3.1.1 Developing Accurate and Explainable Consensus Approaches. Fig 2-Left provides an analysis based on the 3,464 student moderations that were performed on 1,011 resources which also received an instructor evaluation. The figure demonstrates that in cases where instructors provided a rating of 3, 4 or 5, the chance of receiving a true positive (TP) (i.e., students also providing a rating of 3, 4 or 5) is much higher (2399 $\simeq$ 92.7%) than receiving a false negative (FN) (i.e., students providing a rating of 1 or 2) (189 $\simeq$ 7.3%). In contrast, the figure demonstrates that in cases where instructors provided a rating of 1 or 2, the chance of receiving a true negative (TN) (i.e., students providing a rating of 1 or 2) is much lower (162 $\simeq$ 18.5%) than receiving a false positive (FP) (i.e., students providing a rating of 3, 4 or 5) (714 $\simeq$ 81.5%). This demonstrates that only a minority of students seem to be accurately identifying low quality resources. Therefore, simple aggregation functions are likely to perform poorly in learnersourcing systems as they may make a decision based on the judgements from a majority of less

knowledgeable or careless moderators instead of a minority of more knowledgeable or dedicated moderators.

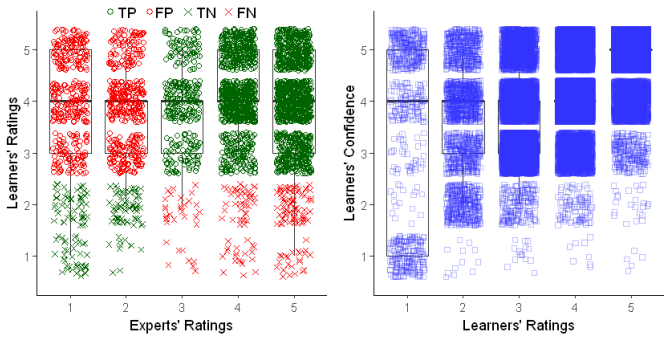


Figure 2: Left: the relationship between student and instructor moderations based on 1,101 resources that were evaluated by both groups; Right: the relationship between students ratings and their provided confidence based on 64,044 moderations.

The consensus algorithm currently used by RiPPLE, as described in [15], employs a set of moderators' ratings of the quality of a resource (1=lowest rating of quality to 5= highest rating of quality) as well as their assessment of their confidence (1 = very low to 5 = very high) in their rating to simultaneously infer the reliability of student moderators and the quality of resources. At a high-level, it follows the principles of the expectation maximisation (EM) algorithm [46]. It first sets the reliability of all student moderators to an initial value of α. In the expectation step, it computes the quality of a resource as a weighted average of the ratings provided by the moderators. In the maximisation step, it updates the rating of the moderators based on the goodness of their decision rating. The update is computed as the height of a Gaussian function based on the distance between the inferred quality rating of the resource and the decision rating provided by a moderator where a smaller distance leads to a greater increase (reward) in the moderator rating. Similarly, a larger distance leads to a greater decrease (punishment) in the moderator rating. A moderator's self-assessment of their confidence in their rating is viewed as the moderator betting on themselves. Selecting a higher confidence level can lead to a bigger reward or punishment, whereas selecting a lower level of confidence can lead to a smaller reward or punishment. The use of this algorithm in a learnersourcing system seemed appropriate because: (1) the computations are light and can easily be performed on the fly; (2) the outcome in terms of inferred rating of the resource and changes in student moderator ratings are easy to explain to the learners; (3) the system encourages moderators to honestly assess their confidence; and (4) the system seemed fair based on our conducted studies with students and instructors.

In practice, the consensus algorithm has generally worked well but has had significant shortcomings. Fig 2-Right demonstrates the relationship between students ratings and their provided confidence based on 64,044 moderations from 2,410 students on 14,559 resources. Overall, it demonstrates that students are most likely to provide a high rating with a high level of confidence. The Dunning-Kruger effect [37], in which low-performing students tend to overestimate their abilities may be one of the factors that contribute to the large number of students who self-assess their confidence as high. Examining the leaderboards for many of the courses that are piloting RiPPLE shows that the students on top of the boards seem to have found a dominant strategy, according to game theory literature [14], to game the system. Given that the vast majority of the resources are approved, an optimal strategy would be to give a rating of 5 and confidence of 5 in the evaluation regardless of the quality of the resource. Our recent work [15], shows that training a machine learning algorithm with simple features extracted from provided comments such as the length and sentiments of the comments to compute the reliability of moderators can significantly increase the correlation between instructor decisions and the inferred decision from the system. However, it is worth noting that addition of simple features such as length of comment in a live setting must be done with care to prevent students from new ways of misusing or gaming the system. For example, a student may figure out by experiment that longer comments tend to increase his reliability rating and therefore submit longer comments that add no value (e.g. submitting "a great great great great great question" instead of "a great question").

3.1.2 Employing Human-in-the-loop and Responsible AI. As discussed in Section 3.1.1, RiPPLE uses a consensus approach to automatically predict the quality of a resource. Predictive models have been extensively used in learning analytics tools and have demonstrated promising results in the automatic identification of students in need of assistance [30, 44]. However, with the wide-adoption of predictive models to support automatic decision-making, there are increasing concerns about using predictive models without human oversight in decision-making tasks that affect individuals [27]. Many factors, such as identifying important variables, dealing with poor quality or imbalanced data, determining the appropriate algorithm and model for the problem at hand, hyperparameter tuning, and knowing when to retrain the algorithm with new data, may bias or reduce the accuracy of the result of a predictive model. To address these issues, human-in-the-loop AI methods that aim to design hybrid system architectures that leverage human intelligence and judgement alongside the power of AI are receiving increasing attention. These methods generally seek human judgement on decisions which the AI model has low decision confidence or where they can help in post-processing the AI-based decision by making sure aspects like fairness are taken into account [70]. In the context of education, the development of fair, accountable and transparent AI systems that rely on human-judgement has been recognised as an important line of research [35, 53, 55]. Within RiPPLE, we have incorporated the following mechanisms for enabling users to provide feedback.

Flagging resources. While interacting with learning resources on the platform, RiPPLE provides students the ability to report resources that have already passed moderation as being incorrect, inappropriate or ineffective. Reported resources are passed on to instructors who can decide whether they need to be deleted, edited or remain as is. To date, 998 resource reports have been submitted by students of which 393 have received an action by an instructor.

Given the relatively large number of reports, we have been considering how we may prioritise and best utilise instructors' availability towards checking resources. An overview of the currently adopted approach is discussed in Section 3.3.

Challenging the outcome of a moderation. In many instances, students who did not agree with the decision made about the quality of a resource, reached out to instructors via email or discussion forums. To streamline the process of enabling students to voice their concern while reducing instructor's workload in responding to enquires, RiPPLE now enables students to provide feedback about moderation decisions through the platform. In the updated workflow, once RiPPLE makes a decision about the quality of a resource, the author and the moderators are notified of the outcome and are given the ability to view the decisions and comments from the moderators. They are then given the ability to provide feedback on whether or not they think the right decision was made. The feedback provided is used in the spot checking algorithm discussed in Section 3.3. Given that this feature has recently been added to the system, we currently do not have data on its use or impact.

Providing ratings, comments and general feedback. The moderation process only captures the evaluations from a few moderators. However, for resources that have gone through moderation and have become available on the platform, students are encouraged to share their opinion in terms of providing ratings or comments on the resources. To date over 188K ratings and 4000 comments have been provided on these resources. RiPPLE also has a general feedback form that enables users to either anonymously or by name provide feedback about the platform. To date we have received 59 feedback comments which have assisted us in making updates to many parts of the platform.

3.1.3 Employing Appropriate Criteria for Evaluating the Quality of Learning Resources. This section reports our data-driven reflections on the significance of the role of the evaluation rubric on students' judgement on the quality of learnersourced content. Fig. 3-left illustrates the initial rubric employed in RiPPLE for evaluation of resources. The top part of the rubric has three Likert scale statements capturing moderators' perceptions on alignment with course content, correctness and coherence of the resource in their evaluation. These three criteria were selected based on suggestions from the literature (e.g., [69]). The bottom part of the rubric has two Likert scale statements capturing moderators' perceptions on the overall quality of the resource (referred to as quality-rating) and their confidence in their judgement as well as space for moderators to provide an open-ended comment. Analysis of 41,048 student moderations based on this rubric revealed the following response distribution on quality-rating: strongly disagree = 1%, disagree = 3%, neutral = 9%, agree = 35%, strongly agree = 51%, where more than a half of the students provided the highest quality-rating to the resources. The length of the provided comments to support quality-rating had a mean of 8.06 words with a standard deviation of 11.12. We were interested in identifying the main criteria that students might have referred to in their justification of their provided quality-rating. Therefore, we conducted a qualitative analysis of 45% (515) of comments obtained from evaluations conducted in the first 8 weeks of a first-year engineering course piloting RiPPLE. To begin the analysis, 5% of the selected comments were manually coded. The codes represented the criteria participants articulated in assessing the effectiveness or quality of the resources they moderated. The codes were used to manually tag the remaining comments. Results revealed references to the three criteria captured by the rubric alignment (10%), correctness (17%) and coherence (20%) as well references to difficulty level (19%) and critical thinking or depth (10%) by students in their comment justifications. Overall, our analysis highlighted the following shortcomings of the rubric being used: (1) it led to the majority of the students rating the quality of resources as the highest possible value which is 5, (2) students provide a very short comment for justifying their quality-rating, and (3) the rubric misses reference to two important criteria (difficulty and critical thinking) that students associate with the quality of a resource.

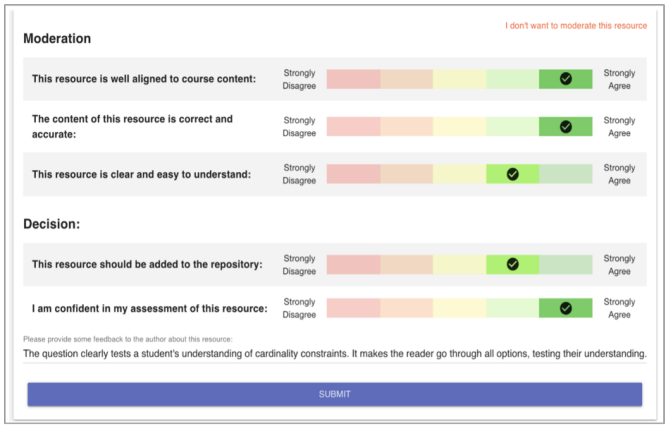
The updated rubric being employed in RiPPLE, illustrated in Fig. 3-Right, attempts to address these limitations. The rubric now has additional criteria that refer to difficulty level and critical thinking. In addition, we have now moved away from Likert scale statements, which are commonly used to capture perceptions in surveys. Instead, we use words that refer to the quality of outcome ranging from poor to outstanding, which is more commonly used in rubrics. Finally, the rubric now specifically ask students to justify their decision and provide feedback rather just having space for a comment without specific instructions. Analysis of 41,048 student moderations based on the new rubric revealed the following response distribution on quality-rating: poor = 1%, needs improvement = 3%, satisfactory = 23%, great = 58%, Outstanding = 15%, which suggests a significant shift in moderators' responses. The length of the provided comments to support quality-rating had a mean of 13.69 with a standard deviation of 16.36.

Comparison of results from the use of the rubrics, as highlighted in Fig. 3, illustrates that the new rubric has managed, to some extent, address the problem of students being "lenient markers", as it has lowered the quality-ratings provided by students. It has also managed to significantly increase the length of the provided comments (justifications), which as discussed in Section 3.1.1, can be employed to increase the accuracy of the system in inferring the quality of resources.

3.2 Incentivising High Quality Contributions

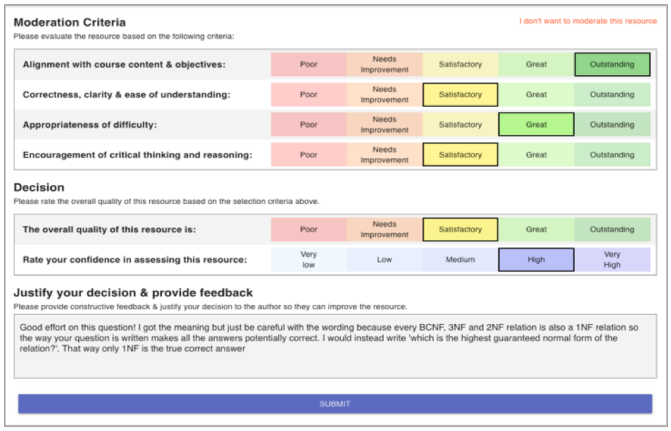
A common phenomenon, referred to as participation inequality or the 90-9-1 rule [49], has been observed in many systems that rely on users to create content. Participation inequality suggests that roughly 90% of users are lurkers (i.e., observe but do not contribute); 9% of users contribute from time to time and 1% of users participate a lot and account for most contributions. While previous work has reported on challenges related to engaging students in learningsourcing [35, 65], implications of participation inequality for learnersourcing systems, and best practices for incentivising a larger portion of students to engage with learnersourcing activities are largely unknown. On a related note, many of the successful systems and platforms that users interact with on a daily basis (e.g., gaming and social media apps or platforms that stream content) have been designed with the prime intention of increasing engagement without considering the quality of the engagement.

The old rubric



# responses	Decision Rating					Comment length
	Strongly Disagree	Disagree	Neutral	Agree	Strongly agree	
41,048	1%	3%	9%	35%	51%	8.06 ± 11.12

The new rubric



# responses	Decision Rating					Comment length
	Poor	Needs Improvement	Satisfactory	Great	Outstanding	
31,012	1%	3%	23%	58%	15%	13.69 ± 16.36

Figure 3: Left: the old rubric; Right: the updated rubric, which is currently being used in RiPPLE.

For example, YouTube would monetarily benefit from a user engaging with a video regardless of whether or not they are paying close attention to the content. In contrast, educational tools have the general intention of improving learning which is more associated with high-quality active engagement rather than purposeless high-quantity engagement. In the context of learnersourcing, past studies have referred to incentivising students to contribute as one of the important learnersoucing challenges that need to be addressed [28, 34, 66]. So how can learnersourcing systems be designed to encourage a large portion of the student population to contribute high-quality learnersourcing activities? In the remainder of this section, we share our experience in developing open learner models for learnersourcing systems (Section 3.2.1), tying learnersourcing to assessment (Section 3.2.2) and employing gamification mechanisms (Section 3.2.3) for attempting to address this challenge.

3.2.1 Open Learner Models for Learnersourcing Systems. Learner models capture an abstract representation of a student's knowledge state. To date, there have been two main use cases for modelling learners: They are (1) employed as a key component of adaptive educational systems to provide personalised feedback or adaptivity functionalities and (2) externalised and made accessible as open learner models (OLMs) [9] to students and instructors with the aim of monitoring, incentivising and regulating learning. In both of these cases, it is essential that the learner model accurately represents the competencies and the current knowledge state of the student. By and large, existing learner models such as Bayesian Knowledge Tracing (BKT) [12], Item Response Theory (IRT) [43] and Elo-based modes [2, 51] are grounded in psychometrics and approximate a student's knowledge state solely based on their performance on assessment items. This can probably be attributed to the fact that in many educational learning systems, students are prominently involved in just answering assessment items. Initially, RiPPLE only leveraged students' responses to assessments items for assessing their mastery level in its open learner model [2]. However, ignoring students learnersourcing contributions in modelling students presented multiple challenges and limitations:

- *Discouraging learnersourcing contributions.* The system, by design, was encouraging (rewarding) students to engage with attempting assessment items while discouraging (ignoring) their learnersourcing contributions.
- *Missing the opportunity to leverage learnersourcing data in modelling students.* Given the strong evidence from the learning sciences that engaging students in higher-order learning tasks such as learnersoucing enhance learning, it is reasonable to expect that leveraging data from students' learnersourcing contributions towards modelling their mastery can improve the accuracy of the model.
- *Advancing the belief that learnersourcing does not contribute to learning.* The explicit association between attempting assessment items and the OLM while ignoring learnersourcing contributions may present the impression that learnersourcing cannot contribute to learning.

So how can we develop OLMs in learnersourcing systems that accurately represent students' mastery level while promoting self-regulation and positive behaviour that contributes to learning in students? We have developed two learner models that leverage data from students' learnersourcing tasks alongside data on attempting assessment items in modelling the knowledge state of learners. The first model extends the knowledge tracing machines (KTMs) framework [60]. At its core, the algorithm utilises the number of creation and evaluation opportunities that the student has had in modelling learners. Results from two empirical studies based on data from past courses that have adopted RiPPLE show that the our proposed model outperform traditional learner models that only use assessment data [1]. Our findings are aligned with results from other recently published papers from the community that suggest

Table 1: Comparison of students' engagement with different activities based two different open learner models

	# Users	# Activities	# Create	# Evaluates	# Answer
Control Group	163	60.17	2.9 ± 1.2	12.9 ± 3.7	44.3 ± 20.6
Experiment Group	161	59.35	3.1 ± 1.3	13.4 ± 4.1	43.3 ± 14.2

modelling knowledge acquisition using student contributions beyond attempting items improves the accuracy of the model [71]. A limitation of this model is that it cannot intuitively be opened to learners to promote self-regulation and positive behaviour.

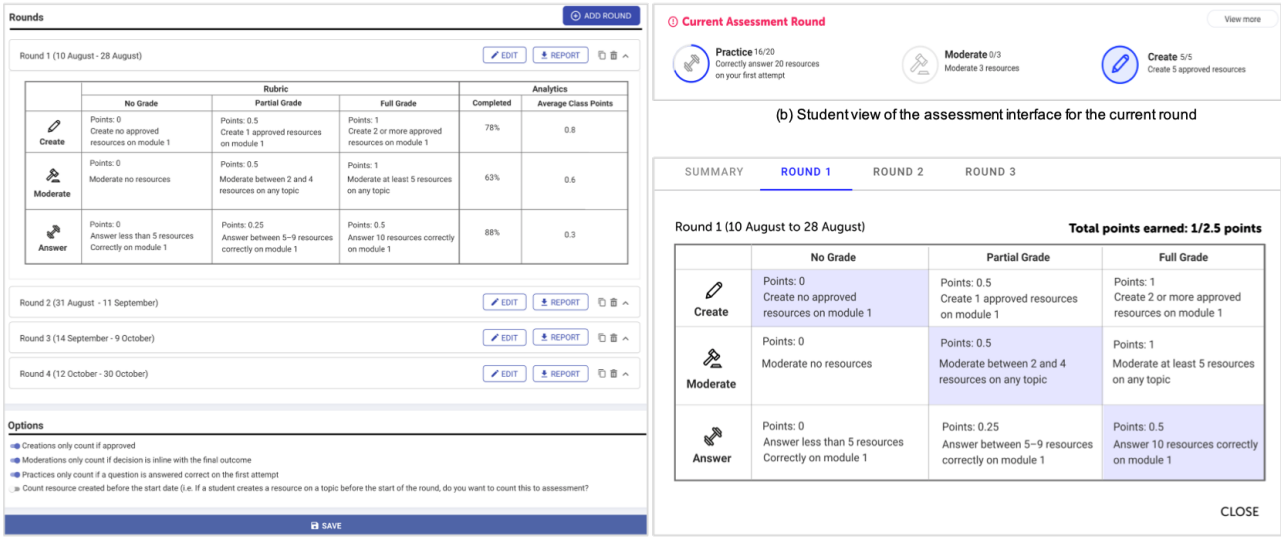
The second proposed model extends the popular Elo model, which has been used as an open learner model in the literature [2]. In the traditional educational Elo approach, students and learning items are viewed as competitors each having a rating representing the mastery level of the student and difficulty level of the item. The probability that the learner answers the item correctly is estimated using a logistic function of a difference between the ratings[51]. In the extended model, learnersourcing contributions are also used for updating the learner model. Similar to the outcome of attempting a resource, outcomes of creating a resource or evaluating a resource are also considered as a win or a loss for the student. Creating a resource that passes moderation is considered a win and not passing moderation is considered a loss. Similarly, a moderation evaluation that agrees with the outcome of the moderation (e.g., a student recommended accepting the resource and the resource was accepted) is considered a win and disagreeing with the outcome is considered a loss. To determine the impact of the Elo-based extension, we conducted a randomised between-subject design experiment in a first-year computer science course with 324 students at The University of Queensland where participants were randomly assigned to one of two open learner models. Participants in the control group had access to the initial OLM used in RiPPLE, which only leveraged students' responses to assessments items for assessing their mastery level. The experiment group had access to the proposed Elo-based extension. Overall, as illustrated in Table 1, the new OLM, had the desirable outcome of increasing students' emphasis on learnersourcing contributions, but the impact was small. Conversely, the adoption of the new model across the platform incentived undesirable behaviour in roughly (1%) of the student population who provided quick and careless evaluations to increase their OLM rating. For example, the student with the highest mastery level in a first-year course in Psychology had submitted over 600 moderations approving all of the evaluated resources with no provided comments.

There are many interesting future directions for further research in development of open learner models for learnersourcing systems. How can we measure the learning acquired by contributing to a learnersoucing task? How can this measure be compared to other tasks such as attempting an assessment item? How can an OLM in a learnersourcing system accurately represent students' mastery level while promoting self-regulation and positive behaviour that contributes to learning in students?

3.2.2 Tying Learnersourcing to Assessment. Initially, RiPPLE was mostly used as a formative tool to support student learning. In this setting, engagement followed the 90-9-1 rule in which the general quality of the contributions was high, but they were mostly coming from a small portion of the student population. With the aim of engaging a larger portion of students in learnersourcing, many courses began tying learnersourcing activities to assessment. The common approach has been introducing three or four rounds where for each round students are asked to create at least one resource and moderate at least five resources. This change introduced two challenges that we have had to address.

(1) *Increasing participation at the cost of decreasing the general quality of engagement.* Tying learnersourcing to assessment has changed the contribution pattern on RiPPLE from a 90-9-1 discussed above to 25-50-25 where roughly 25% of the students provide high quality contributions and exceed assessment expectations, 50% of the students aim to just satisfy the assessment requirements with minimal effort and 25% of the students entirely disengage with the activity. We have attempted introducing quality measures to increase the quality of contributions. For example, the assessment criteria ask for creation of an effective resources, where a resource is considered effective if it passes the moderation process, which is administered by students and the teaching team. For moderations, introducing quality measures has been more challenging. In two offerings, one in psychology and one computer science, the assessment criteria was defined as having effective moderations, where an effective moderation was defined as a moderation that agrees with the inferred outcome (approved or rejected) by the system. However, this method of defining an effective moderation had the undesirable effect of encouraging the minority of critical reviewers who were helping to identify low-quality resources to follow the majority and provide high ratings to maximise their chance of agreeing with the inferred outcome. Another computer science course defined an effective moderation as "completing the moderation rubric and providing a detailed justification for your judgement as well as constructive feedback on how the resource can be improved. Simply saying a resource is "good" does not qualify". This method did not have the undesirable effect of encouraging the minority of critical reviewers to become less critical. However, an interesting observation from this course was that students sometimes were writing longer generic comments (e.g., "Good explanation and question. I think this will work well.", "It's simple and straightforward which is useful for platforms like RiPPLE." or "Great question, the solution was well explained and simple to understand."), which were not very helpful as they did not relate to the actual content of the resources. In all cases, we found that the quality control may require some spot-checking from the teaching team, as discussed in Section 3.3, to examine the quality of the contributions.

(2) *Supporting Assessment logistics.* The second problem of tying learnersourcing to assessment was related to the logistics of enabling instructors to communicate the requirements of the assessment and for students to track their progress towards the completion of the assignment. There were many enquires from students related to not knowing whether or not they have met the requirements for completion of their assessment. Figure 4 provides an overview of how we have attempted to support the use of assessment in RiPPLE. Figure 4-a shows the instructor view where they can create a rubric for an assessment round. The creation of the rubric is scaffolded using multiple steps were instructors determine



(a) Instructor view of the assessment interface

(b) Student view of the assessment interface for the current round

(c) Student view of grades for the completed rounds

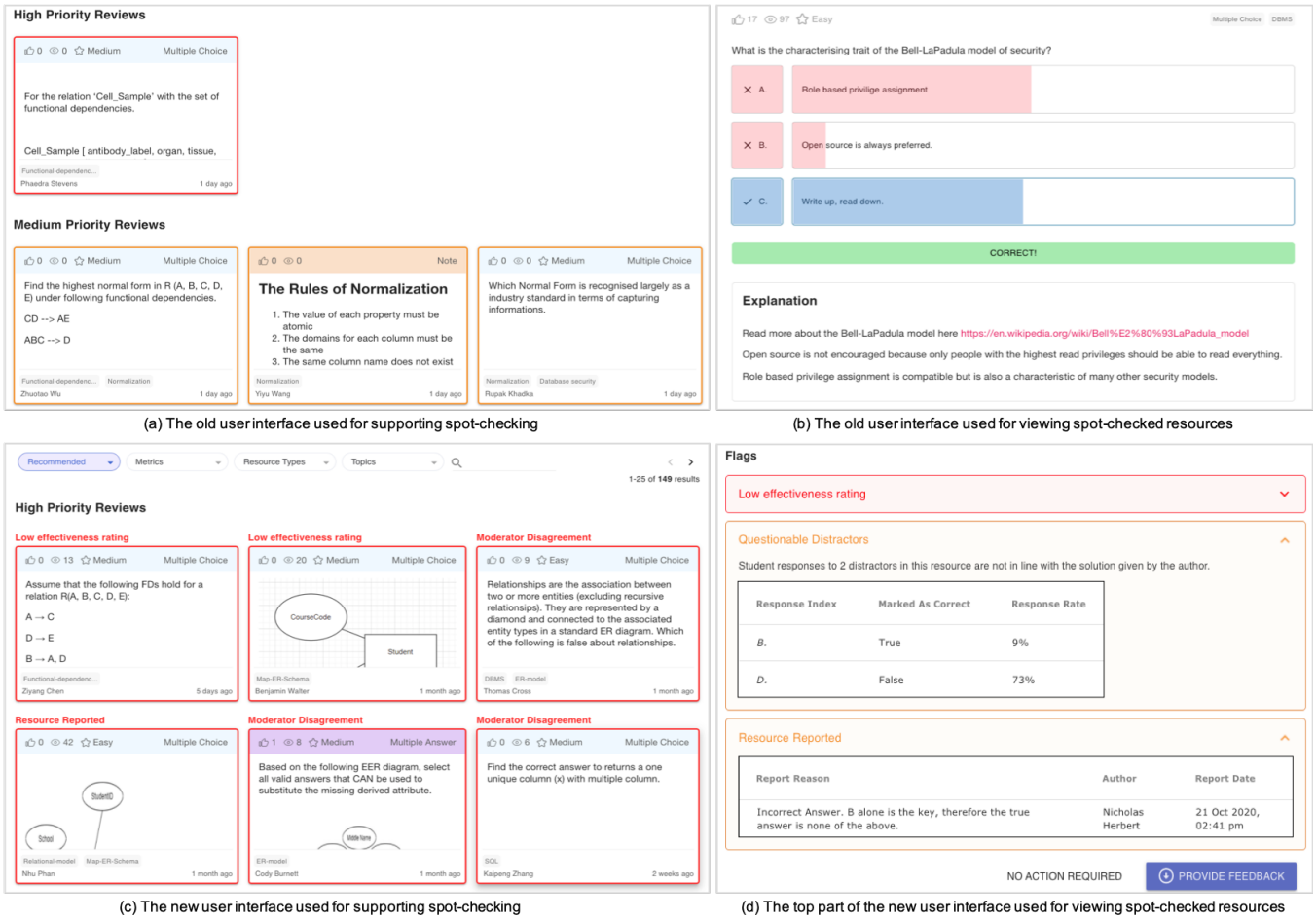
Figure 4: An overview of assessment support in RiPPLE.

their assessment requirements. At any time, instructors can view general statistics about completion rate and average grade per criteria or simply download grades in CSV format. Figure 4-b shows the student view where students can track their progress on the current round. Clicking on the "view more" button shows them their grades for the completed rounds as shown in Figure 4-c.

3.2.3 Employing Gamification Mechanisms. There is a general consensus that motivation is regarded as one of the most important factors leading to academic success [41]. Employing mechanisms that are typically used in games in non-game contexts, commonly referred to as gamification, has been viewed as a viable option to increase participation and engagement in many different settings including education. While there have been contradictory findings on the effects of interacting with gamified systems in education [57], evidence suggests that if game elements target behaviours that can improve learning then gamification can have a positive impact on student engagement and learning [18]. Data from a survey on RiPPLE conducted in a graduate course with 75 students on database principles at The University of Queensland illustrates students' view on gamification features of RiPPLE. A total of 56 students completed the survey, which had 15 questions that were based on 5-point Likert scale statements. 73% of respondents agreed or strongly agreed that the weekly *awards* motivated them to use RiPPLE and 78% of respondents agreed or strongly agreed that the *leaderboard* motivated them to use RiPPLE. Future work aims to formally investigate the impact of these different gamification strategies on student engagement and performance using the methodology used in [18].

3.3 Empowering Instructors with Actionable and Explainable Insights

In recent years, the learning analytics community has developed many tools and technologies that provide a large range of analytics to help instructors make data-informed decisions about students' learning. For example, learning analytics dashboards have been successful in developing visualisations that help with sensemaking and displaying information about student performance and behaviour. In spite of some promising results, the actual impact of learning analytics dashboards has been found to be relatively low, questioning their ability in presenting feedback that can meaningfully be translated into actionable recommendations to improve learning [8, 45]. On the other hand, a diverse range of educational tools have utilised recommender and adaptive engines to personalise the user experience of students or instructors [21]. Commonly, these systems operate as a "black-box", giving users no insight into the rationale of their recommendations, which can lead to trust issues, especially when users do not agree with the proposed recommendations [29, 59]. The Human-Computer Interaction research community has been looking at how users interact with recommender systems and how such interaction can feed back and be leveraged to improve system design. Classic studies [13] looked at the impact of the recommender system interface on how the system is perceived by users. They observed how the rating scale and the display of prediction have the strongest impact on how users perceive system effectiveness. Their conclusions suggest a users' preference towards finer-grained rating scales and a strong sensitiveness to recommendation inaccuracies. [42] looked at how to improve the usability and user acceptance of recommender systems by combining automated algorithms with the ability for users to interact and explore the collection of items showing the improvements obtained by such a hybrid approach. To understand the importance of explainable systems, [39] recently looked at how recommender systems can explain the reasons why they suggest items as compared to humans. Authors concluded that the quality of explanations provided by humans is still perceived as superior and they observed that the better the explanation the better the recommendation is perceived to be by users. Some initial work on



(a) The old user interface used for supporting spot-checking

(b) The old user interface used for viewing spot-checked resources

(c) The new user interface used for supporting spot-checking

(d) The top part of the new user interface used for viewing spot-checked resources

Figure 5: An overview of the old and new approaches of supporting spot-checking.

the development of transparent recommender systems have been conducted in the learning analytics community [3], but overall, the topic is still under-developed and under-researched and.

So how can learnersourcing systems couple analytics and recommendations to empower instructors with actionable and explainable insights to guide student learning? Below, we provide an example of coupling analytics and recommendations with the aim of empowering instructors with actionable and explainable insights from RiPPLE.

Given the limited availability of instructors, RiPPLE incorporates a spot-checking algorithm [62] to identify resources that would benefit the most from being reviewed by an expert. RiPPLE's spot-checking approach has focused on identifying resources that have passed moderation but are likely to be incorrect or ineffective. At a high-level, the spot-checking algorithm in RiPPLE employs a range of human-driven metrics (e.g., high-disagreement in moderation evaluations, a high ratio of downvotes in comparison to upvotes) and data-driven metrics (e.g., assessments items that have a low discrimination index or questionable distractors where the popular answer is not the one proposed by the author) to categorise resources into having high, medium, low or no priority for being reviewed. Figure 5-a illustrates the old version of the user interface used for supporting spot-checking, where instructors were able to see the priority category of each resource without any justification. Figure 5-b illustrates the old user interface of what instructors would have seen once they clicked on a recommended resource, which is essentially the page presenting the resource. While instructors expressed a general appreciation for the spot-checking algorithm, many of them were not sure why a resource was selected as having a high priority for being reviewed.

Figure 5-c illustrates the updated interface where we have attempted to provide rationale for the spot-checking recommendations. There are two main additions: (1) resources are now tagged with metrics that have been employed in the categorisation of their priority class and (2) instructors are provided with a search bar where they can provide additional constraints such as use of particular metrics or topics for selecting resources to review. Figure 5-d illustrates the top part of the updated interface that instructors now see when they click on a recommendation. The main addition is that instructors are now provided with additional rationale for why a resource was flagged for being reviewed. In the given example,

rationale for why the resource was flagged based on three metrics are provided: (1) low effectiveness as the resource has received more down-votes than up-votes, (2)stats on questionable distractors that seem to be incorrect, and (3) an overview of a student report indicating the question is incorrect. Below the flag component, instructors can view the actual resource as illustrated in Figure 5-b. Once an instructor has reviewed the provided rationale and the resource, they can either clear the proposed flags or take action by deleting the resource or providing feedback to the author. Informal feedback has been very positive about this change. Experiments on the impact of providing these additional analytics to support recommendations on instructors' perceptions and behaviour are underway.

4 CONCLUSION

The overarching aim of this paper is to contribute to the growing literature on the development of effective learnersourcing systems and more broadly technological educational solutions that support learner-centred education at scale. Data-driven reflections and lessons learned throughout the paper can be summarised into the following suggestions for developing learnersourcing systems: (1) employ accurate and explainable consensus approaches for assessing the quality of resources (see Section 3.1.1), (2) empower students and instructors to raise concerns in relation to irresponsible use of AI and fairness (see Section 3.1.2), (3) reflect on the criteria used in the evaluation of the quality of resources and investigate its impact (see Section 3.1.3), (4) employ OLMs that accurately show student mastery while promoting high-quality learnersourcing contributions (see Section 3.2.1), (5) develop required features and logistics to support tying learnersourcing to assessment (see Section 3.2.2), (6) employ various gamification mechanisms to incentivise high-quality learnersourcing contributions (see Section 3.2.3), (7) couple analytics and recommendations to empower instructors with actionable and explainable insights (see Section 3.3) and finally (8) conduct rigorous empirical studies to investigate the impact of various features of the platform. In terms of implications for practice and research, our findings reiterate the emphasis on the interdisciplinary nature to enhance the impact of learning analytics. In particular, the findings are inline with the definition of learning analytics as crossroads between data science, design, and educational theory [24]. The findings specifically emphasize that while the adoption of machine learning in learning analytics can be beneficial, it is not sufficient to achieve desirable outcomes. Instead, a careful integration of machine with pedagogical interventions and design of user interfaces is essential for learning (e.g., for learner modelling and moderation of student evaluations). Likewise, design of user interfaces and pedagogical interventions in learning analytics should harness the benefits that stem from developments in artificial intelligence (e.g., to improve peer feedback or moderation). It is also critical to approach iteratively the development and implementation of systems empowered with learning analytics. Each iteration should evaluate effectiveness of interventions introduced including those based on pedagogy, user interface design, machine learning, and their combinations.

Acknowledgements. This work is partially supported by the ARC Discovery Project (Grant No. DP190102141).

REFERENCES

[1] Solmaz Abdi, Hassan Khosravi, and Shazia Sadiq. 2020. Modelling Learners in Crowdsourcing Educational Systems. In *Artificial Intelligence in Education*, Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.). Springer International Publishing, Cham, 3–9.

[2] 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.

[3] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Complementing Educational Recommender Systems with Open Learner Models. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 360–365.

[4] Philip Adey. 1999. The Science of Thinking, and Science for Thinking: A Description of Cognitive Acceleration through Science Education (CASE). Innodata Monographs 2. (1999).

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

[6] 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.

[7] Sameer Bhatnagar, Amal Zouaq, Michel C Desmarais, and Elizabeth Charles. 2020. Learnersourcing Quality Assessment of Explanations for Peer Instruction. In *European Conference on Technology Enhanced Learning*. Springer, 144–157.

[8] Robert Bodily and Katrien Verbert. 2017. Trends and issues in student-facing learning analytics reporting systems research. In *Proceedings of the seventh international learning analytics & knowledge conference*. 309–318.

[9] Susan Bull. 2020. There Are Open Learner Models About! *IEEE Transactions on Learning Technologies* 13 (2020), 425 – 448.

[10] Susan Bull, Blandine Ginon, Clelia Boscolo, and Matthew Johnson. 2016. Introduction of learning visualisations and metacognitive support in a persuadable open learner model. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 30–39.

[11] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 169–178.

[12] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[13] D Cosley, SK Lam, I Albert, JA Konstan, and J Riedl. 2003. Is Seeing Believing? Now Recommender Interfaces Affect Users' Opinions. In *CHI 2003*.

[14] Jacques Crémer and Richard P McLean. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society* (1988), 1247–1257.

[15] Ali Darvishi, Hassan Khosravi, and Shazia Sadiq. 2020. Utilising Learnersourcing to Inform Design Loop Adaptivity. In *Addressing Global Challenges and Quality Education*, Carlos Alario-Hoyos, María Jesús Rodríguez-Triana, Maren Scheffel, Inmaculada Arnedillo-Sánchez, and Sebastian Maximilian Dennerlein (Eds.). Springer International Publishing, Cham, 332–346.

[16] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*. 469–478.

[17] 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*. 51–58.

[18] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 311.

[19] Shayan Doroudi, Ece Kamar, and Emma Brunskill. 2019. Not Everyone Writes Good Examples but Good Examples Can Come from Anywhere. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 12–21.

[20] Shayan Doroudi, Joseph Williams, Juho Kim, Thanaporn Patikorn, Korinn Ostrow, Douglas Selent, Neil T Heffernan, Thomas Hills, and Carolyn Rosé. 2018. Crowdsourcing and Education: Towards a Theory and Praxis of Learnersourcing. International Society of the Learning Sciences, Inc.[ISLS].

[21] Hendrik Drachsler, Katrien Verbert, Olga C Santos, and Nikos Manouselis. 2015. Panorama of recommender systems to support learning. In *Recommender systems handbook*. Springer, 421–451.

[22] Carol Evans. 2013. Making sense of assessment feedback in higher education. *Review of educational research* 83, 1 (2013), 70–120.

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

[24] Dragan Gašević, Vitomir Kovanović, and Srećko Joksimović. 2017. Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice* 3, 1 (2017), 63–78.

[25] Elena L. Glassman, Aaron Lin, Carrie J. Cai, and Robert C. Miller. 2016. Learnersourcing Personalized Hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. 1626–1636.

[26] Philip J. Guo, Julia M. Markel, and Xiong Zhang. 2020. Learnersourcing at Scale to Overcome Expert Blind Spots for Introductory Programming: A Three-Year Deployment Study on the Python Tutor Website. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. 301–304.

[27] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *ACM SIGKDD ICKDDM*. 2125–2126.

[28] 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.

[29] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[30] Ya-Han Hu, Chia-Lun Lo, and Sheng-Pao Shih. 2014. Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior* 36 (2014), 469–478.

[31] Evgeny Karataev and Vladimir Zadorozhny. 2016. Adaptive social learning based on crowdsourcing. *IEEE Transactions on Learning Technologies* 10, 2 (2016), 128–139.

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

[33] Hassan Khosravi, George Gyamfi, Barbara E. Hanna, and Jason Lodge. 2020. Fostering and Supporting Empirical Research on Evaluative Judgement via a Crowdsourced Adaptive Learning System. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 83–88.

[34] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics* 6, 3 (2019), 91–105. https://doi.org/10.18608/jla.2019.63.12

[35] Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Development and Adoption of an Adaptive Learning System: Reflections and Lessons Learned. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 58–64.

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

[37] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[38] James A Kulik and JD Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.

[39] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[40] Jihyun Lee and Hyoseon Choi. 2017. What affects learner's higher-order thinking in technology-enhanced learning environments? The effects of learner factors. *Computers & Education* 115 (2017), 143–152.

[41] Elizabeth A Linnenbrink and Paul R Pintrich. 2002. Motivation as an enabler for academic success. *School psychology review* 31, 3 (2002), 313–328.

[42] Benedikt Loepp, Tim Hussein, and Jüergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3085–3094.

[43] Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

[44] Farshid Marbouti, Heidi A Diefes-Dux, and Krishna Madhavan. 2016. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education* (2016), 1–15.

[45] W. Matcha, N. A. Uzir, D. Gašević, and A. Pardo. 2020. A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies* 13, 2 (2020), 226–245.

[46] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, 6 (1996), 47–60.

[47] Steven Moore, Huy A Nguyen, and John Stamper. 2020. Evaluating Crowdsourcing and Topic Modeling in Generating Knowledge Components from Explanations. In *International Conference on Artificial Intelligence in Education*. Springer, 398–410.

[48] Catherine Mulryan-Kyne. 2010. Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education* 15, 2 (2010), 175–185.

[49] Jakob Nielsen. 2006. Participation inequality: Encouraging more users to contribute. *www. useit. com/alertbox/participation_inequality. html* (2006).

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

[51] Radek Pelánek, Jan Papoušek, Jiří Řihák, Vít Stanislav, and Juraj Nižnan. 2017. Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction* 27, 1 (2017), 89–118.

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

[53] Vanessa Putnam and Cristina Conati. 2019. Exploring the Need for Explainable Artificial Intelligence in Intelligent Tutoring Systems.. In *IUI Workshops*.

[54] Shiva Shabaninejad, Hassan Khosravi, Marta Indulska, Aneesha Bakharia, and Pedro Isaias. 2020. Automated Insightful Drill-down Recommendations for Learning Analytics Dashboards. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/3375462.3375539

[55] Simon J Buckingham Shum and Rosemary Luckin. 2019. Learning analytics and AI: Politics, pedagogy and practices. *British journal of educational technology* 50, 6 (2019), 2785–2793.

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

[57] Rob van Roy and Bieke Zaman. 2018. Need-supporting gamification in education: An assessment of motivational effects over time. *Computers & Education* 127 (2018), 283–297.

[58] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4 (2011), 197–221.

[59] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 351–362.

[60] Jill-Jênn Vie and Hisashi Kashima. 2019. Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 750–757.

[61] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[62] Wanyuan Wang, Bo An, and Yichuan Jiang. 2018. Optimal Spot-Checking for Improving Evaluation Accuracy of Peer Grading Systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[63] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. 1–10.

[64] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 405–416.

[65] Daniel S Weld, Eytan Adar, Lydia B Chilton, Raphael Hoffmann, Eric Horvitz, Mitchell Koch, James A Landay, Christopher H Lin, and Mausam Mausam. 2012. Personalized Online Education-A Crowdsourcing Challenge.. In *HCOMP@ AAAI*. Citeseer.

[66] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. 379–388.

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

[68] Naomi Winstone and David Carless. 2019. *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

[69] Min Yuan and Mimi Recker. 2015. Not all rubrics are equal: A review of rubrics for evaluating the quality of open educational resources. *International Review of Research in Open and Distributed Learning* 16, 5 (2015), 16–38.

[70] Fabio Massimo Zanzotto. 2019. Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.

[71] Siqian Zhao, Chunpai Wang, and Shaghayegh Sahebi. 2020. Modeling Knowledge Acquisition from Multiple Learning Resource Types. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*. 313–324.

[72] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.