# Utilising Learnersourcing to Inform Design Loop Adaptivity

Ali Darvishi, Hassan Khosravi(✉), and Shazia Sadiq

School of Information Technology and Electrical Engineering,
The University of Queensland, St Lucia, QLD 4072, Australia
{a.darvishi,h.khosravi}@uq.edu.au, Shazia@itee.uq.edu.au

**Abstract.** Design-loop adaptivity refers to data-driven decisions that inform the design of learning materials to improve learning for student populations within adaptive educational systems (AES). Commonly in AESs, decisions on the quality of learning material are based on students' performance, i.e., whether engaging with the material led to learning gains. This paper investigates an alternative approach for design adaptivity, which utilises students' subjective ratings and comments to infer the quality of the learning material. This approach is in line with the recent shift towards learner-centred learning and learnersourcing, that aim to transform the role of students from passive recipients of content to active participants that engage with various higher-order learning tasks including evaluating the quality of resources. In this paper, we present a suite of aggregation-based and reliability-based methods that can be used to infer the quality of learning material based on student ratings and comments. We investigate the feasibility and accuracy of the methods in a live learnersourcing educational platform called RiPPLE that provides the capacity to capture subjective ratings and comments from students. Empirical data from the use of RiPPLE in a first-year course on information systems are used to evaluate the presented methods. Results indicate that the use of a combination of reliability-based methods provides an acceptable level of accuracy in determining the quality of learning resources.

**Keywords:** Adaptive educational systems · Learnersourcing · Crowdsourcing in education

## 1 Introduction

Adaptive educational systems (AESs) [4] make use of data about students, learning process, and learning products to provide an efficient, effective and customised learning experience for students by dynamically adapting learning content to suit their individual abilities or preference. Adaptation to the needs of an individual or an entire student population can be guided via the following three adaptation loops, namely design-loop, task-loop and step-loop adaptivity [4]. Design-loop adaptivity refers to data-driven decisions to update learning


C. Alario-Hoyos et al. (Eds.): EC-TEL 2020, LNCS 12315, pp. 332–346, 2020.
https://doi.org/10.1007/978-3-030-57717-9_24

material to improve learning for the entire student population. In contrast, task-loop and step-loop adaptivity refer to data-driven decisions the system makes to select instructional tasks or actions within a task for an individual learner. Commonly, adaptation in all three loops is guided based on students' performance [4]. As an example, data on the extent to which engagement with a resource leads to learning gains for the student population can be used to infer the quality of resources as part of the design loop.

In a recent trend, researchers from a diverse range of fields (e.g., Learning at Scale (L@S), Artificial Intelligence in Education (AIED), Computer Supported Cooperative Work (CSCW), Human-Computer Interaction (HCI) and Educational Data Mining (EDM)) have explored the possibility of employing crowdsourcing approaches to support high-quality, learner-centred learning at scale. The use of crowdsourcing in education, often referred to as learnersourcing, is defined as "a form of crowdsourcing in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves" [16]. Recent progress in the field highlights the potential benefits of employing learnersourcing, and the rich data collected through it, towards addressing the challenges of delivering high quality learning at scale. In particular, With the increased enrolments in higher education, educational researchers and educators are beginning to use learnersourcing in novel ways to improve student learning and engagement [3,7,8,10,11,15,25–27].

Following this trend, this paper aims to investigate whether learnersourcing can be used as an effective mechanism to inform design loop adaptivity in AESs. We present a suite of methods that can take subjective ratings and comments about the quality of a learning resource from students to infer its true quality. The presented methods are categorised into two groups: aggregation-based methods and reliability-based methods. Aggregation-based methods rely on statistical aggregations such as mean and median. Reliability-based methods aim to infer the reliability of each student so that more reliable students could have a larger contribution towards the computation of the final decision. The reliability-based methods presented include a method that uses the submitted subjective ratings, a method that uses the text of the provided comments and another method that considers the alignment between the two numeric and linguistic ratings.

To contextualise the problem under investigation within an educational setting, we present an AES called RiPPLE that relies on learnersourcing for design loop adaptivity. Empirical data from the use of RiPPLE in a first-year computer science course at The University of Queensland is used to compare and contrast the suite of the presented methods. Results suggest that traditional reliability-based inference methods that have been demonstrated to work effectively in the context of other crowdsourcing systems may not work well in education. A potential explanation is that crowdsourcing systems generally rely on the wisdom of the majority; however, the majority of the crowd may not necessarily be wise in the educational domain. Our findings further suggest that using multiple reliability-based methods in conjunction may be an effective way to improve the results.

In the remaining paper, we first present related work on learnersourcing and a summary of quality control and consensus approaches used in learnersourcing. Our problem is formalised in Sect. 3, where we provide the suite of methods proposed for our study. Section 4 presents the RiPPLE system and provides details of how the learnersourced ratings and comments are collected in RiPPLE. The evaluation of the proposed methods is presented in Sect. 5, and the paper is concluded in Sect. 6 with a summary of contributions and limitations.

## 2 Related Work

The use of learningsourcing is inspired by contemporary models of learning that have emphasised the importance of learner-centred approaches that engage learners in higher-order learning activities, which enable learners to develop their own vision, reasoning, and judgement to extend understanding, including lifelong learning [5,13]. There are many successful examples of learnersourcing systems. For example [7] empowers learners to author multiple-choice questions, AXIS [26] uses students to generate, revise, and evaluate explanations as learners solve problems, UpGrade [25] sources student open-ended solutions to create scalable learning opportunities, RiPPLE [14] learnersources generation of learning activities which are used as part of an adaptive educational system. Another popular use of crowdsourcing in education is peer grading. There are many successful examples of peer grading systems including Mechanical TA [29], Peer Assessment [22], PeerGrade [28], Aropa [20] and PeerScholar [19].

Whereas these studies and supporting tools have made significant contribution to enhancing student experience, peer learning and improvement in learning outcomes, currently there is limited understanding of how learnersourcing can be used effectively for improving the design of learning resources. One of the main challenges in this regard is assessing the quality of learning resources created or evaluated by learners in contrast to experts. Due to the potential that the decision made by an individual learner might be incorrect, many learnersourced evaluation systems employ a redundancy-based strategy and assign the same tasks to multiple learners. 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 [30]. Many of the state-of-the-art crowd consensus approaches rely on machine learning algorithms (e.g., [23]) to simultaneously infer the true outcome and workers' reliability. While using machine learning algorithms have significantly improved the accuracy of the models compared to averaging aggregation functions, these methods often lack understandability and transparency (in terms of how individuals were rated and how a final decision was made). The use of black-box outcomes seems to be particularly inadequate for educational settings and where educators strive to provide extensive feedback to enable learners to develop their own vision, reasoning, and appreciation for inquiry and investigation and fairness. As such, many such systems have focused on simple aggregation-based methods such as use of mean and median [7,19,28] which are easy to

understand an transparent in terms of decision making process; however, they lack the required accuracy.

Much of the existing work on the need for open and transparent models in education has been conducted in the field of open learner models [6] where models are often opened through visualisations, as an important means of supporting learning in various applications such as learning analytics dashboards [6], assessment tools [12], intelligent tutoring systems [9], educational recommender systems [1], and adaptive learning platforms [14]. The available literature on crowd consensus approaches and open learner models indicates that the development of crowd consensus approaches that can be used in learner-centred learning which are accurate but also explainable and fair are still under-developed and –investigated.

## 3 Problem Definition and Inference Models

In what follows, Sect. 3.1 presents a formal definition of the problem under investigation. Section 3.2 presents three traditional aggregation-based inference models for inferring the quality of a learning resource. Finally, Sect. 3.3 presents three reliability-based inference models for inferring the quality of a learning resource. Table 1 provides a summary of the notation used within this section.

### 3.1 Problem Definition

Let's assume that $U_N = \{u_1 \dots u_N\}$ denotes a set of students who are enrolled in a course in an educational system, where $u_i$ refers to an arbitrary student. Let $Q_M = \{q_1 \dots q_M\}$ present the content model, denoting a repository of learning resources that are available to students within the system, where $q_j$ refers to an arbitrary learning resource. Furthermore, let $D_{N \times M}$ denote decision ratings where $1 \leq d_{ij} \leq 5$ shows the decision rating given by user $u_i$ to resource $q_j$. let $C_{N \times M}$ denote comments that are provided to accompany decision ratings where $c_{ij}$ denote the comment provided by user $u_i$ on resource $q_j$. Using the information available in $D_{N \times M}$ and $C_{N \times M}$, our aim is to infer $\hat{R_M} = \{\hat{r_1} \dots \hat{r_M}\}$, where $1 \leq \hat{r_j} \leq 5$ shows the quality of $q_j$.

### 3.2 Aggregation-Based Inference Models

A widely used method for inferring an outcome from a set of individual decisions is to use statistical aggregations such as mean or median. We will also present a third example that uses aggregation functions to identify and address user bias. In the explanation of the models given in the remainder of this section, we will assume that decision ratings and associated comments from a set of users $\{u_1 \dots u_k\}$ on a resource $q_j$ are used to infer $\hat{r_j}$.

**Table 1.** Notation used in the problem definition and the presented approaches.

| | |
|---|---|
| Input parameters | |
| $U_N$ | A set of students $\{u_1 \dots u_N\}$ who are enrolled in the course |
| $Q_M$ | A repository of learning resources $\{q_1 \dots q_M\}$ available within the system |
| $D_{N \times M}$ | A two dimensional array in which $1 \leq d_{ij} \leq 5$ shows the decision rating given by user $u_i$ to resource $q_j$ |
| $C_{N \times M}$ | A two dimensional array in which $c_{ij}$ denote the comment provided by user $u_i$ on resource $q_j$ |
| Aggregation-based models | |
| $B_N$ | A set of users' bias $\{b_1 \dots b_N\}$ in which $b_i$ shows the bias of student $u_i$ in rating the quality of resources |
| $\bar{d}_i$ | The average decision rating of user $u_i$ |
| $\bar{d}$ | The average decision rating across all users |
| Reliability-based models | |
| $W_N$ | A set of users' reliability $\{w_1 \dots w_N\}$ in which $w_i$ infers the reliability of a user $u_i$ |
| $\alpha$ | The initial value of the reliability of all students |
| $LC_{N \times M}$ | A two dimensional array in which $lc_{ij}$ denote the length of the comment provided by user $u_i$ on resource $q_j$ |
| $F^R_{N \times M}$ | A function where $f^R_{ij}$ determines the quality of the rating provided by $u_i$ for $q_j$ |
| $F^L_{N \times M}$ | A function where $f^L_{ij}$ approximates the 'effort' of $u_i$ in evaluating $q_j$ |
| $F^A_{N \times M}$ | A function where $f^A_{ij}$ approximates the alignment between the rating and comment provided by $u_i$ on $q_j$ |
| Output | |
| $\hat{R_M}$ | A set of $M$ ratings $\{\hat{r_1} \dots \hat{r_M}\}$ where each rating $1 \leq \hat{r_j} \leq 5$ shows the quality of resource $q_j$ |

*Mean.* A simple solution is to use mean aggregation, where for $\hat{r_j} = \frac{\sum_{i=1}^{k} d_{ij}}{k}$. There are two main drawbacks to using mean aggregation: (1) it is strongly affected by outliers and (2) it assumes that the contribution of each student has the same quality, whereas in reality, students' academic ability and reliability may vary quite significantly across a cohort.

*Median.* An alternative simple solution is to use $\hat{r_j} = Median(u_1, \dots u_k)$. A benefit of using median is that it is not strongly affected by outliers; however, similar to mean aggregate, it assumes that the contribution of each student has the same quality, which is a strong and inaccurate assumption.

*User Bias.* Some students may consistently underestimate (or overestimate) the quality of resources. We introduce the notation of $B_N$, where $b_i$ shows the bias

of user $u_i$ in rating. Introducing a bias parameter has been demonstrated to be an effective way of handling user bias in different domains such as recommender systems and crowdconsensus approaches [17]. We first compute $\bar{d_i}$ as the average decision rating of a user $u_i$. We then compute $\bar{d} = \frac{\sum_{i=1}^{N} \bar{d_i}}{N}$ as the average decision rating across all users. The bias term for user $u_i$ is computed as $b_i = \bar{d_i} - \bar{d}$. A positive $b_i$ shows that $u_i$ provides higher decision ratings compared to the rest of the cohort and similarly a negative $b_i$ shows that $u_i$ provides lower decision ratings compared to the rest of the cohort. To adjust for bias, the quality of resource $q_j$ can be inferred as $\hat{r_j} = \frac{\sum_{i=1}^{k}(d_{ij} - b_i)}{k}$.

### 3.3 Reliability-Based Inference Models

Students within a cohort can have a large range of academic abilities. We introduce the notating of $W_N$, where $w_i$ infers the reliability of a user $u_i$ so that more reliable students could have a larger contribution towards the computation of the final decision. Many methods have been introduced in the literature for computing reliability of users [30]. The problems of inferring the reliability of users $W_M$ and quality of resources $R_M$ can be seen as solving a "chicken-and-egg" problem where inferring one set of parameters depend on the other. If the true reliability of students $W_M$ were known then an optimal weighting of their decisions could be used to estimate $R_M$. Similarly, if the true quality of resources $R_M$ were known, then the reliability of each student $W_N$ could be estimated. In the absence of ground truth for either, we present three heuristic methods that can easily be embedded in live educational systems where students can view updates to their reliability score. In all three examples we (1) set the reliability of all students to an initial value of $\alpha$; (2) compute $\hat{r_j}$ for a resource $q_j$ based on current values of $w_1, \ldots w_k$ and $d_1, \ldots d_k$ and $c_1, \ldots c_k$; (3) update $w_1, \ldots w_k$. The methods of computing $\hat{r_j}$ and updating $w_1, \ldots w_k$ in each of the three methods are described below.

*Rating.* In this method, the current ratings of the users and their given decisions are utilised for computing the quality of the resources and reliabilities. In this method, $\hat{r_j}$ and $w_i$ are computed using Formula 1 as follows:

$$\hat{r_j} = \frac{\sum_{i=1}^{k} w_i \times d_{ij}}{\sum_{i=1}^{k} w_i}, w_i := w_i + f_{ij}^R \tag{1}$$

where $F_{N \times M}^R$ is a function in which $f_{ij}^R$ determines the 'goodness' of $d_{ij}$ based on $\hat{r_j}$ using the distance between the two $dif_{ij} = |d_{ij}$ - $\hat{r_j}|$. Formally, $f_{ij}^R$ is computed as the height of a Gaussian function at value $dif_{ij}$ with centre 0 using $f_{ij}^R = \delta \times \frac{e^{-(dif_{ij})^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} - \frac{\delta}{2}$ where the hyper-parameters $\sigma$ and $\delta$ can be learned via cross-validation. Informally, $f_{ij}^R$ provides a large positive value (reward) in cases where $dif_{ij}$ is small and it provides a large negative value (punishment) in cases where $dif_{ij}$ is large.

*Length of Comment.* The reliability of a user decision in the previous scenario relies on the numeric ratings provided for a resource and it does not take into account how much effort was applied by a user in the evaluation of a resource. In this method, the current ratings as well as decisions and comments of users are utilised for computing the quality of the resources and updating reliabilities. We introduce the notation of $LC_{N\times M}$, where $lc_{ij}$ shows the length of comments (i.e., number of words) provided by user $u_i$ on resource $q_j$. $\hat{r_j}$ and $w_i$ are computed using Formula 2 as follows:

$$\hat{r_j} = \frac{\sum_{i=1}^{k}(w_i + f_{ij}^L) \times d_{ij}}{\sum_{i=1}^{k}(w_i + f_{ij}^L)}, w_i := w_i + f_{ij}^L \tag{2}$$

where $F_{N\times M}^L$ is a function in which $f_{ij}^L$ approximates the 'effort' of $u_i$ in answering $q_j$ based on the length of comment $lc_{ij}$. Formally, $f_{ij}^L$ is computed based on the logistic function $\frac{c}{1+ae^{-k\times lc_{ij}}}$ where the hyper-parameters $c$, $a$ and $k$ of the logistic function can be learned via cross-validation. Informally, $f_{ij}^L$ rewards students that have provided a longer explanation for their rating and punishes students that have provided a shorter explanation for their rating.

*Rating-Comment Alignment.* The previous two reliability-based models take into account the similarity of the students' numeric rating with their peers and the amount of effort they have spent on moderation by the length of their comments. Here, the alignment between the ratings and comments provided by a user are considered. In this method, $\hat{r_j}$ and $w_i$ are computed using Formula 3 as follows:

$$\hat{r_j} = \frac{\sum_{i=1}^{k}(w_i + f_{ij}^A) \times d_{ij}}{\sum_{i=1}^{k}(w_i + f_{ij}^A)}, w_i := w_i + f_{ij}^A \tag{3}$$

Where $F_{N\times M}^A$ is a function where $f_{ij}^A$ approximate the alignment of the rating $d_{ij}$ and the comment $c_{ij}$ a user $u_i$ has provided for a resources $q_j$. A sentiment analysis tool that assesses the linguistic features in the comments provided by the students on each resource is used to classify the words in terms of emotions into positive, negative and neutral. The Jockers-Rinker sentiment lexicon provided in the SentimentR package is applied here to compute a sentiment score between $-1$ to 1 with 0.1 interval. This package assigns polarity to words in strings with valence shifters [18,21]. For example, it would recognize this sample comment "This question is Not useful for this course" as negative rather than indicating the word "useful" as positive.

*Combining Reliability Functions.* Any combination of the presented three reliability functions can also be considered. For example, Formula 4 uses all three of the rating, length and alignment methods for reliability.

$$\hat{r_j} = \frac{\sum_{i=1}^{k}(w_i + F_{ij}^L + F_{ij}^A) \times d_{ij}}{\sum_{i=1}^{k}(w_i + F_{ij}^L + F_{ij}^A)}, w_i := w_i + F_{ij}^R + F_{ij}^L + F_{ij}^A \tag{4}$$

## 4 The RiPPLE Platform

At its core, RiPPLE is an adaptive educational system that dynamically adjusts the level or type of instruction based on individual student abilities or preferences to provide a customised learning experience [14]. Figure 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 an Elo-based rating system for approximating the knowledge state of users [2] with the results translated into coloured bars. The lower part of the RiPPLE screen displays learning resources recommended to a student based on his/her learning needs using the recommender system outlined in [14].

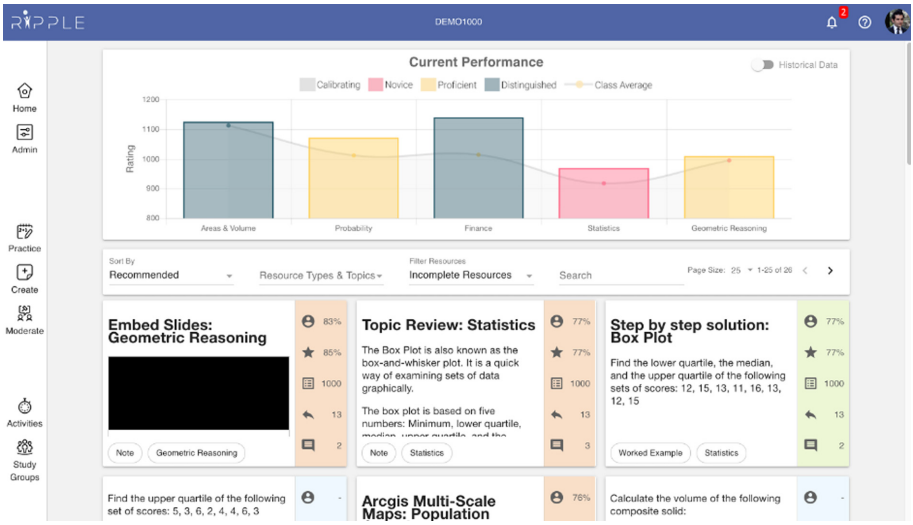


**Fig. 1.** Overview of student modelling and recommendation page of RiPPLE (Color figure online)

*Learnersourcing.* To provide customised learning for students with different knowledge states, adaptive educational systems require large repositories of learning resources, which are commonly created by domain experts [4]. Such systems are therefore expensive to develop and challenging to scale. Instead of relying on domain experts as developers, RiPPLE uses a learner-sourcing approach to engage students in the creation, moderation and evaluation of learning resources (activities). This does not only reduce the cost of content generation, it also holds the potential to foster students' higher-order skills. However, as

students are developing their expertise, it is likely that some of the learning resources created are ineffective, inappropriate or incorrect. 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, thus requiring them to judge the quality of their peers' work.

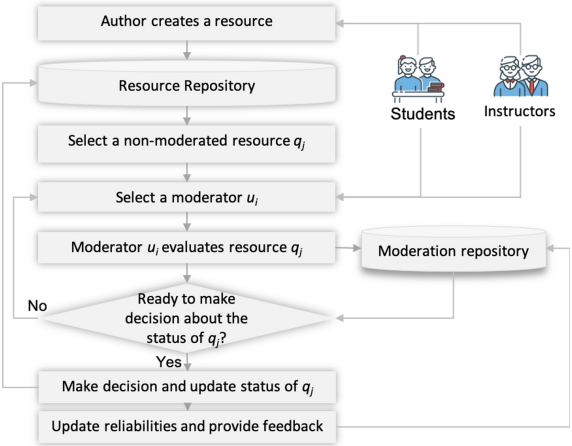


**Fig. 2.** Overview of the moderation process in RiPPLE

Figure 2 provides an overview of the moderation process in RiPPLE. Both students and instructors can author learning resources in RiPPLE. Resources authored by instructors are automatically added to the existing pool of the learning resources which are available to all enrolled users. Resources authored by students will go through a formal moderation process where students and instructors judge the quality of the resource. For a non-moderated resource $q_j$, ripple assigns a moderator $u_i$ from the pool of available moderators to evaluate it. Once the evaluation is complete, RiPPLE determines whether or not the resource needs to be evaluated by further moderators. RiPPLE provides two options for how this decision is to be made: (1) instructors determine the number of student moderations required per resource and (2) RiPPLE determines whether it can confidently make a judgement based on the current pool of available evaluations or an expert opinion or further moderations are required. In this option, at a high level of generality, RiPPLE considers the level of agreement between moderators; if there is a strong agreement, then it will make a decision based on the formed consensus. Otherwise, it will request an instructor to evaluate the resource or seek further evaluations. In both cases, RiPPLE uses spot checking algorithms [24] to present resource that would benefit the most from expert judgement to instructors. Moderations from instructors are considered final, meaning their decisions are considered as the ground truth without considering evaluations from students. Once RiPPLE is ready to make a decision, it will update the

status of $q_i$ to be approved and added to the pool of the available resources or to be denied and removed from the pool. In both cases, the reliability of the moderators are updated using the rating algorithm presented in Sect. 3.3 and feedback about the outcome is provided to the author, moderators, and instructors.

**Moderation Feedback**

| Moderator | Decision | Rating | Confidence | Weight | Comment |
|---|---|---|---|---|---|
| | 3 | 1037 | 4 | 30% | Domain constraint is a type of Integrity constraint. |
| | 3 | 1032 | 5 | 37% | Question should read: "Which of the following does NOT violate domain constraints. Also the use of "integrity constraints" in the question is misleading as it implies that domain constraints are not a type of integrity constraint. Also the wording of the third line of the question could be slightly better worded. If these issues are resolved, then the question would be a very good resource to help understand domain constraints. |
| | 1 | 1144 | 4 | 33% | Should the question not be "which DOES NOT violate" ? Since it states which does, but then lists the only correct instance as the answer? |

**Result: Denied (2.3)**

**Fig. 3.** An example of the moderation outcome and feedback provided by RiPPLE.

Figure 3 demonstrates an example of how moderation outcome and feedback are shared with instructors. Instructors can view the name [removed in the figure], decision, current rating, confidence level (as determined by the moderator), the weight of contribution towards making the final decision, and comments provided by each moderator. The author and moderators can see the decision, confidence level, contribution weight and the provided comment; however, they cannot view the identity or the current rating of the moderator.

## 5 Evaluation

In this section, we use empirical data from the use of RiPPLE in a first-year course on information systems to evaluate the suite of the presented methods[1]. Section 5.1 presents the data set used for the evaluation, Sect. 5.2 presents the metric used in the evaluation and Sect. 5.3 presents the results.

### 5.1 Data Set

The data set used in this study is obtained from piloting RiPPLE during the first six weeks of a course on information systems at The University of Queensland.

[1] Approval from our Human Research Ethics Committee at The University of Queensland was received for conducting this evaluation #2018000125.

A total of 353 students who were enrolled in the course have submitted 2,722 moderations on 611 learning Resources that were created by the students themselves. From these resources, 96 of them have also received moderations from an instructor which are put aside as the test set. These 96 resources have received a total of 373 moderations from 192 students. The remaining 515 resources were used for the training set. These resources have received a total of 2,349 moderations from 347 students.

Figure 4 provides further information about the training data set. Figure 4(a) shows the total number of moderations, Fig. 4(b) shows the average rating across all moderations, and Fig. 4(c) shows the average length of comments in words across all moderations. This figure demonstrates that students have quite diverse behaviour in terms of their moderations with the majority moderating between 1–16 resources with a mean of $6.8 \pm 5.3$ moderations, having an average rating between 2.7–5 with a mean of $4.1 \pm 0.6$ and writing comments with 0–44 words with a mean of $17.4 \pm 15.3$ words. This figure also shows that resources have received diverse moderations with the majority receiving 1–10 moderations with a mean of $4.6 \pm 1.9$, having an average rating between 2.6–5 with a mean of $4.1 \pm 0.6$ and having comments with 1–36 words with a mean of $15.6 \pm 9.3$ words.

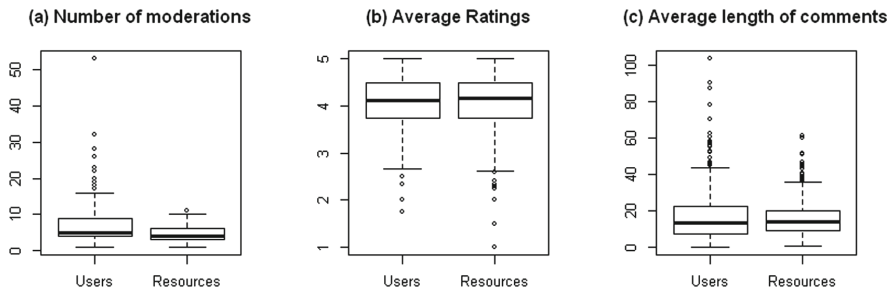


**Fig. 4.** Visualisations of the training data set

The test set mostly includes resources in which making a decision without expert judgement on quality was challenging, which resulted in the resource being presented to instructors for moderation. For this data set, ratings of the instructors have a mean of $2.83 \pm 1.30$ and ratings from students have a mean of $3.50 \pm 1.17$

### 5.2 Evaluation Metric

To evaluate the presented algorithms, we compute the correlation between domain expert ratings and student ratings based on the provided ratings. We report the $r - value$ and $p - value$ of the regressed model where $r - value$ is the Pearson correlation coefficient and $p - value$ is the two-sided $p - value$ obtained from a Wald test for which the null hypothesis is that the slope of the regressed

line is zero. The $r - value$ represents the strength of the relationship, and the $p - value$ determine the statistical significance of the result.

### 5.3 Results

Table 2 shows the result of comparisons between the instructor and student ratings based on the suite of the presented inference models. In all ten cases, the results are statistically significant with $p < .001$.

**Table 2.** Correlation between the instructor and student ratings in different models

| Type | Method | Correlation | |
|---|---|---|---|
| | | r-value | p-value |
| Aggregation-based methods | Mean | 0.52 | p < 001 |
| | Median | 0.51 | p < 001 |
| | User Bias | 0.53 | p < 001 |
| Reliability-based methods | $F^R$ | 0.54 | p < 001 |
| | $F^L$ | 0.60 | p < 001 |
| | $F^A$ | 0.62 | p < 001 |

| Type | Method | Correlation | |
|---|---|---|---|
| | | r-value | p-value |
| Combinations | $F^{R+L}$ | 0.59 | p < 001 |
| | $F^{R+A}$ | 0.60 | p < 001 |
| | $F^{L+A}$ | 0.60 | p < 001 |
| | $F^{R+L+A}$ | **0.63** | p < 001 |

Aggregation-based methods such as mean and median rely only on the current subjective ratings for each resource regardless of users' history data or any other measure of the users' performance (that is quality of the rating). The mean and median models achieve a correlation of 0.52 and 0.51, respectively. The user bias model considers the overall rating history of users as an adjustment to the mean method and achieves a correlation of 0.53. Unsurprisingly, we observe in our evaluations that the outcomes of the quality rating using aggregation-based methods have the lowest correlations with the expert decisions.

Reliability-based models infer the reliability of a user. The performance of these models is evidently better than what is obtained using aggregation based methods. Here, users' reliability scores are computed by three different measures: (1) $F^R$ numeric rating of students compared to their peers for the resource at hand, (2) $F^L$ length of comment provided by the user and (3) $F^A$ alignment between rating and comment on the resource indicated by sentiment analysis.

$F^R$, a weighted averaging method that only uses students numeric ratings, shows little improvement at 0.54 correlation compared to the user bias model. This result was surprising as methods that infer reliability based on rating have been demonstrated to work well in many systems that crowdsource decision-making [30]. In the absence of ground truth, an assumption made by this model is that the system can rely on the wisdom of the majority in decision making. However, the system would fail to perform well in cases where the majority is not wise. As an example, if the majority poorly evaluates a resource, then the weighted average would be closer to those that made a poor judgement. This leads to rewarding poor evaluators and punishing reliable ones. In contrast, $F^L$

relies on those that have made a more significant effort rather than relying on the majority. The use of this model has led to a notable improvement of performance to 0.60 correlation. Results from these two models suggest that there is a minority of students in the used data set that have put a more significant effort in moderating which are more reliable than the majority. $F^A$, which relies on the alignment of numeric and linguistic ratings is the most effective single model which has a 0.62 correlation.

We also consider and report the performance of the four possible combinations of the reliability-based methods. The first two rows, i.e. $F^{R+L}$ and $F^{R+A}$, demonstrate that adding the length of comments and the rating-comment alignment to the numeric subjective rating results in an improvement of performance from what can be achieved via $F^R$. Similarly, $F^{L+A}$ achieves better performance than what can be achieved via $F^L$ individually. However, the best outcome of combinations is gained using $F^{R+L+A}$, which demonstrates that considering the wisdom of the majority in combination with recognition of effort as well as alignment of numeric and linguistic rating would achieve the greatest improvement with a 0.63 correlation in performance.

## 6 Conclusion and Future Work

Learnersourcing is an emerging area of interest for adaptive educational systems. Engaging learners in the creation and evaluation of learning resources has been shown to have beneficial outcomes for student experience and learning gains, while overcoming issues of scale and timely feedback. However a key challenge in this regard is how to determine the quality of the contributions made by learners in contrast to experts. In this paper, we address this challenge by focussing on the moderation process in building learning resource repositories. We present a number of aggregation and reliability based methods for assessing the quality of learnersourced numeric and linguistic ratings. Our results indicate that reliability-based methods that consider the wisdom of the majority in combination with recognition of efforts as well as alignment of numeric and linguistic ratings perform the best in terms of accuracy of the quality judgement. We posit that using our proposed method, resource moderation can be undertaken without the need for expert intervention.

While simple feature extractions based on the length and sentiments of the comments have significantly improved the accuracy of the quality judgement based on an offline data set, simple addition of these reliability-based features into a live system may promote misuse via gaming the system (e.g., submitting a long comment that repeats one word). Future work focuses on employing more advanced artificial intelligence and natural language processing methods for judging the quality of moderations based on comments.

## References


1. Abdi, S., Khosravi, H., Sadiq, S., Gasevic, D.: Complementing educational recommender systems with open learner models. In: Proceedings of the Tenth International Conference LAK, pp. 360–365 (2020)
2. Abdi, S., Khosravi, H., Sadiq, S., Gasevic, D.: A multivariate Elo-based learner model for adaptive educational systems. In: Proceedings of the Educational Data Mining Conference, pp. 462–467 (2019)
3. Alenezi, H.S., Faisal, M.H.: Utilizing crowdsourcing and machine learning in education: literature review. Educ. Inf. Technol. **25**, 1–16 (2020)
4. Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R.: Instruction based on adaptive learning technologies. In: Mayer, R.E., Alexander, P. (eds.) Handbook of Research on Learning and Instruction. Routledge, New York (2016)
5. Boud, D., Soler, R.: Sustainable assessment revisited. Assess. Eval. High. Educ. **41**(3), 400–413 (2016)
6. Bull, S., Ginon, B., Boscolo, C., Johnson, M.: Introduction of learning visualisations and metacognitive support in a persuadable open learner model. In: Proceedings of the 6th Conference on Learning Analytics & knowledge, pp. 30–39 (2016)
7. Denny, P., Hamer, J., Luxton-Reilly, A., Purchase, H.: Peerwise: students sharing their multiple choice questions. In: Proceedings of the Fourth International Workshop on Computing Education Research, pp. 51–58 (2008)
8. Doroudi, S., et al.: Crowdsourcing and Education: Towards a Theory and Praxis of Learnersourcing. International Society of the Learning Sciences, London (2018)
9. Guerra, J., Hosseini, R., Somyurek, S., Brusilovsky, P.: An intelligent interface for learning content: combining an open learner model and social comparison to support self-regulated learning and engagement. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 152–163 (2016)
10. Heffernan, N.T., et al.: The future of adaptive learning: does the crowd hold the key? Int. J. Artif. Intell. Educ. **26**(2), 615–644 (2016)
11. Karataev, E., Zadorozhny, V.: Adaptive social learning based on crowdsourcing. IEEE Trans. Learn. Technol. **10**(2), 128–139 (2016)
12. Khosravi, H., Cooper, K.: Topic dependency models: graph-based visual analytics for communicating assessment data. J. Learn. Anal. **5**(3), 136–153 (2018)
13. Khosravi, H., Gyamfi, G., Hanna, B.E., Lodge, J.: 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, pp. 83–88 (2020)
14. Khosravi, H., Kitto, K., Joseph, W.: RiPPLE: a crowdsourced adaptive platform for recommendation of learning activities. J. Learn. Anal. **6**(3), 91–105 (2019)
15. Kim, J., Nguyen, P.T., Weir, S., Guo, P.J., Miller, R.C., Gajos, K.Z.: Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 4017–4026 (2014)
16. Kim, J., et al.: Learnersourcing: improving learning with collective learner activity. Ph.D. thesis, Massachusetts Institute of Technology (2015)
17. Krishnan, S., Patel, J., Franklin, M.J., Goldberg, K.: A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In: Proceedings of the 8th Conference on Recommender systems, pp. 137–144 (2014)
18. Naldi, M.: A review of sentiment computation methods with R packages. arXiv preprint arXiv:1901.08319 (2019)

19. Paré, D.E., Joordens, S.: Peering into large lectures: examining peer and expert mark agreement using peerscholar, an online peer assessment tool. J. Comput. Assist. Learn. **24**(6), 526–540 (2008)
20. Purchase, H., Hamer, J.: Peer-review in practice: eight years of Aropä. Assess. Eval. High. Educ. **43**(7), 1146–1165 (2018)
21. Rinker, T.: Sentimentr: calculate text polarity sentiment, version 2.4. 0 (2018)
22. Shnayder, V., Parkes, D.C.: Practical peer prediction for peer assessment. In: Fourth AAAI Conference on Human Computation and Crowdsourcing (2016)
23. Venanzi, M., Guiver, J., Kazai, G., Kohli, P., Shokouhi, M.: Community-based bayesian aggregation models for crowdsourcing. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 155–164 (2014)
24. Wang, W., An, B., Jiang, Y.: Optimal spot-checking for improving evaluation accuracy of peer grading systems. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Wang, X., Talluri, S.T., Rose, C., Koedinger, K.: Upgrade: sourcing student open-ended solutions to create scalable learning opportunities. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, pp. 1–10 (2019)
26. Williams, J.J., et al.: Axis: generating explanations at scale with learnersourcing and machine learning. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale, pp. 379–388 (2016)
27. Willis, A., Davis, G., Ruan, S., Manoharan, L., Landay, J., Brunskill, E.: Key phrase extraction for generating educational question-answer pairs. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, pp. 1–10 (2019)
28. Wind, D.K., Jørgensen, R.M., Hansen, S.L.: Peer feedback with peergrade. In: ICEL 2018 13th International Conference on e-Learning, p. 184. Academic Conferences and publishing limited (2018)
29. Wright, J.R., Thornton, C., Leyton-Brown, K.: Mechanical TA: partially automated high-stakes peer grading. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 96–101 (2015)
30. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: is the problem solved? Proc. VLDB Endowment **10**(5), 541–552 (2017)