Evaluating the Quality of Learning Resources: A Learnersourcing Approach

Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Gianluca Demartini

***Abstract*—Learnersourcing is emerging as a viable approach for mobilizing the learner community and harnessing the intelligence of learners as creators of learning resources. Previous works have demonstrated that the quality of resources developed by students is quite diverse with some resources meeting rigorous judgmental criteria while other resources are ineffective, inappropriate, or incorrect. Consequently, to effectively utilize these large repositories of resources in student learning, there is a need for a selection and moderation process to separate high-quality resources from low-quality ones in such repositories. Instructors and domain experts are potentially the most reliable source for doing this task; however, their availability is often quite limited. This paper explores whether and how learnersourcing, as an alternative approach, can be used for evaluating the quality of learning resources. To do so, we first follow a data-driven approach to explore students' ability in judging the quality of learning resources. Results from this study suggest that, overall, ratings provided by students strongly correlate with ratings from experts; however, students' ability in evaluating learning resources can also vary significantly. We then present a consensus approach based on matrix factorization (MF) and indicate how it can be used for improving the accuracy of aggregating learnersourced decisions. We also demonstrate how utilizing information on student performance and incorporating ratings from domain experts on a limited number of learning resources can be leveraged to further improve the accuracy of the results.**

***Index Terms*—Consensus algorithm, contributing student pedagogy, crowdsourcing in education, learnersourcing, matrix factorization.**

I. Introduction

THE concept of learnersourcing refers to a form of crowdsourcing with students as a crowd, in which "students collectively contribute novel content for future students while engaging in a meaningful learning experience themselves" [1]. Learnersourcing has been mainly inspired by the success of crowdsourcing, which has proved itself as an effective problem-solving paradigm in several areas. However, the fundamental difference between the two is that unlike crowdsourcing that outsources the tasks to an undefined crowd through an open call, learnersourcing relies on students as a specialized crowd who are naturally motivated and engaged in their learning [1]. In addition, while crowdsourcing leverages the crowd for primarily just getting the job done, learnersourcing aims to introduce respectful, mutually beneficial learning partnerships between students (as experts in training) and academics (as experts) [1]. Successful examples of learnersourced artifacts created by students include traces of videos watched by students [2], annotating videos for future students [1], creating open-ended artifact, such as solutions [3], or explanations [4], [5], and curriculum design [6].

One important approach in which learnersourcing has been used in education is to invite students to create repositories of learning resources that can be leveraged to create novel learning experiences. This approach, as Hill deliberates [7], originates from the student as producer model [8] which itself builds on the existing literature on student-centered learning [9], inquiry-based learning [10], and contributing student pedagogy [11]. The use of learnersourcing for the task of content creation is associated with two types of benefits. The first benefit is associated with transforming the role of students from passive recipients of content to active creators of course material [12]. Based on Bloom's taxonomy of the cognitive model, creating learning resources by students engages them in the highest order of learning [13]. Previous studies have reported that placing the responsibility of content creation in the hands of students reinforces and deepens students' understanding of the course content through engaging them in cognitively demanding tasks [14], [15]. This, in turn, highlights the significance of representing students' work in a clear and logical fashion [16], encourages them to reflect on the course objectives [12], enhances their conceptual understanding [17], and facilitates the capacity for students to relate their learning to their personal experiences, which is the core principle of constructivist theory [18], [19]. The second benefit comes from harnessing the creative power of the students towards the development of large repositories of learning resources [20]. Availability of large repositories of high-quality learning resources can provide great benefits in different contexts. For example, it can be used by students for studying [21], by instructors for creating exams or assignments [20], and by adaptive learning systems for recommending personalized learning resources [22]. Previous studies have shown that students can create repositories of high-quality learning resources that meet rigorous judgmental criteria [16], [20], [23]. In fact, students as the authors of learning resources may have an advantage over instructors, since resources developed by students may have a lower chance of suffering from an expert's blind spot [22].

Despite the advantages mentioned above, learnersourcing with students raises a potential risk that the created content may be ineffective, inappropriate or incorrect [22]. As such, there is a need for a selection and moderation process that

Manuscript received August 27, 2019; revised July 24, 2020 and November 22, 2020; accepted February 6, 2021. Date of publication TBD, date of current version November 22, 2020. (*Corresponding author*: Solmaz Abdi.)

The authors are with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia (e-mail: solmaz.abdi@uq.edu.au; h.khosravi@uq.edu.au; shazia@itee.uq.edu.au; g.demartini@uq.edu.au).

will separate high-quality resources from low-quality ones in such repositories. In education, it is a common practice to have domain experts evaluate the quality of learning resources [20], [24]; however, the large size of repositories of learning resources created by students makes this approach an expensive and time-consuming activity for domain experts. A potential solution would be to use the learnersourcing approaches that take students' evaluation of the quality of created learning resources into account for inferring the right quality of resources. Not only this is an affordable approach, but also encourages students to think critically and analytically about learning resources, reinforces their reflection on their own created content, and motivates discussions around the desired learning outcomes [21]. Furthermore, it has the potential to help with developing the evaluative judgment of students, which is an important aspect of the learning process [25]–[27].

The goal of the presented article is to explore whether and how learnersourcing can be used for evaluating the quality of learning resources. The utilization of students as the evaluators of learning resources relies on the assumption that students, as non-experts, have the ability to evaluate the contributions of their peers [11]. While extensive research effort has been put on studying the ability of students in developing learning resources (e.g., [1], [7], [21]), empirical investigation of the plausibility of involving students in evaluating the quality of available learning resources in educational systems, learnersourced by their peers, has received little attention. To fill this research gap, we follow a data-driven approach and present a study in which the evaluations provided by students to the quality of learning resources are compared with the evaluations provided by domain experts. Results from this study suggest that generally, students' evaluations are strongly correlated with evaluations from experts, and students with a stronger academic performance may be able to evaluate more accurately.

We then present a second study that focuses on how learnersourced evaluations can be integrated towards estimating the quality of learning resources. In decision-making tasks, due to the potential that the decision made by an individual might be incorrect, it is common to employ a redundancy-based strategy, assign the same tasks to multiple users, and then utilize an appropriate consensus mechanism for optimal integration towards making an accurate final decision. In the context of students as evaluators, this would entail having multiple students evaluating the quality of the same learning resource. While designing mechanisms for accurate integration of crowdsourced data through machine learning algorithms has been studied extensively within the crowdsourcing community [28], the use of these approaches for aggregating learnersourced decisions has received little attention. Accordingly, in this study, we present a consensus approach based on matrix factorization (MF) and explore its accuracy for aggregating learnersourced evaluations. We also explore how the availability of additional information on student performance and incorporation of evaluations from domain experts on a limited number of learning resources can be utilized to further increase the accuracy of the results. The data and source code to reproduce the results of both of these case studies are available through a GitHub repository [29].

The rest of this paper is organized as follows: Section II reviews the background on learnersourcing and consensus approaches. Section III presents a data-driven approach to study the plausibility of determining the quality of learnersourced learning resources using the evaluations provided by students. Section IV presents the second study and introduces a machine learning algorithm for aggregating learnersourced evaluations. Section V discusses the findings of the two conducted studies and presents the concluding remarks and plans for the future.

II. Background

We discuss related works in two different realms; (1) prior work related to learnersourcing and particularly its application for content creation, and (2) prior work related to consensus approaches on improving the accuracy of information integration in the tasks performed by a group of non-experts.

A. Learnersourcing

Earlier works on students' contributions in the development of learning resources root from inquiry-based learning [10]. Barak *et al.* [14] conducted a study on an MBA course, in which students enrolled in the course were responsible for contributing learning resources to an online repository as well as ranking the resources contributed by other students to the repository. His findings indicated that students who actively contributed resources to the online repository and assessed other students' contributed resources performed better in their final examination compared to other students. Yu *et al.* [30] reported on a web-based online learning system which enables students to contribute learning resources to the repository of the system. Before the final submission of the created resources to the repository of the system, they are reviewed by other students in terms of their quality. Their findings revealed that high-performing students tended to contribute more learning resources, while low-performing students tended to generate fewer resources. Later works on this topic were introduced under the umbrella title of contributing students pedagogy (CSP) [11]. For example, Denny *et al.* [16], [21] developed one of the most successful web-based learning systems called PeerWise that uses students' generated learning resources in the form of multiple-choice questions (MCQs) to build the repositories of learning resources. Students can also rate the quality of learning resources developed by other students. PeerWise has been deployed in a wide range of undergraduate courses around the world with detailed analyses of student engagement and performance [31]. The results of these analyses indicate that students who contribute to the creation of resources use higher order of thinking skills to generate good quality resources [32] as well as taking an active role in their learning [33], and that there is a positive association between students' engagement in PeerWise and their academic achievements [34].

More recently, as crowdsourcing has become a topic of interest within Learning at Scale (L@S) and Artificial Intelligence in Education (AIED) communities, works on this

topic are widely introduced under the title of learnersourcing. For example, Hill [7] presented a conceptual framework for learnersourcing content creation in education. The findings of a conducted case study using this framework in an undergraduate psychology course suggested that engagement in such activities helps students to construct the relationship between their previous knowledge and experiences and the content of the course which, in turn, improves their ability in solving more complex problems. Heffernan *et al.* [5] proposed using learnersourcing as part of the popular ASSISTments platform, Williams *et al.* [4] presented an Adaptive eXplanation Improvement System (AXIS) that uses learnersourcing to generate, revise, and evaluate explanations as learners solve problems. Farasat *et al.* [35] developed Crowdlearning in which students collaboratively create learning resources for each other, and Karataev *et al.* [36] proposed a framework that combines concepts of crowdsourcing, online social networks, and adaptive systems to provide personalized learning pathways for students. Khosravi *et al.* [22] reported on a learnersourcing adaptive learning platform that recommends personalized learning activities to students from a pool of learnersourced learning resources that are generated by educators and the students themselves.

While many of the discussed studies have relied on students for determining the quality of peer-created learning resources (e.g., [14], [21], [22], [30]), the investigation of students' ability in judging the quality of learning resources was not conducted by any of them. Accordingly, in this paper, we investigate the plausibility of using students' evaluations of the quality of learning resources for judging their quality.

B. Consensus Approaches

With the widespread adoption of crowdsourcing in the problems that require human intelligence, optimal integration of crowdsourced decisions in the absence of ground truth has become a challenging task [37]. Traditional approaches for crowdsourcing consensus use general statistical aggregations, such as the arithmetic mean, median, or majority vote [38]. However, previous studies showed that the quality of the crowdsourced data varies across different crowdsourcers and is affected by several factors, such as skill, underlying motivation, and level of expertise [37], [39], [40]. Therefore, designing mechanisms for improving the efficiency and accuracy of the integration of crowdsourced decisions has received attention within the crowdsourcing community.

One of the classical crowdsourcing consensus approaches uses validated gold standard data to identify good crowdsourcers from bad with the assumption that crowdsourcers who provide incorrect answers to the gold standard questions can be omitted from further evaluations [41], [42]. An important limitation of this approach is that the gold standard data is not always easily accessible [43]. Another common crowdsourcing consensus approach relies on machine learning algorithms to aggregate crowdsourced data. Probabilistic models and in particular Expectation-Maximization (EM), which are generally developed in the context of categorization tasks, such as [37], [44]–[48], are well-known examples of these class of approaches. Another classic examples of machine learning-based approaches are the ones that regard the aggregation problem as an information recommendation problem and utilize collaborative filtering recommendation techniques, such as item-based collaborative filtering, MF, and tensor factorization, to aggregate crowdsourced data [40], [49]–[51]. Despite the success of aggregation mechanisms based on machine learning in crowdsourcing context, in the educational setting, prior works that rely on learners to evaluate the quality of peer-created artifacts generally employ statistical aggregations, such as averaging aggregation for integrating students' decisions. For example, many peer grading systems, such as Mechanical TA [52], Peer Assessment [53], Peergrade [54], Aropä [55], and peerScholar [56], use statistical aggregations.

Inspired by the success of machine learning approaches for the integration of crowdsourced decisions, in this paper, we propose using a machine learning algorithm for performing crowd consensus on learnersourced evaluations. We employ MF in our presented approach because of the following three reasons: (1) Evaluating learning resources by students requires effort from students, and since it is a voluntary task, each student might evaluate a limited number of resources. Consequently, the collected ratings for learning resources are generally sparse and imbalanced, with the consensus evaluation for each learning resource is determined by only a few students. This is similar to the context of product review or movie rating that the availability of reviews from each user is quite sparse. MF has a proven capacity to directly alleviate the issue of sparsity in the collected ratings from different users [49]. MF can induce a latent feature vector for each student and each learning resource which can be used for inferring student ratings on all learning resources [40], [49]. (2) MF can work in both unsupervised or semi-supervised settings, which enables us to explore the accuracy of the algorithm with and without the presence of information from experts, and (3) MF works well in the presence of auxiliary data [57], which enables us to explore the impact of having access to additional information, such as student performance.

III. Study 1: Students as Evaluators of Peer-Created Learning Resources

This study aims to investigate the plausibility of determining the quality of learnersourced learning resources using the evaluations provided by students to their quality. To do so, we follow a data-driven approach and compare the ratings provided by students to the quality of a set of learning resources to the ratings provided by domain experts on the same set of learning resources as the gold standard.

A. Method

The research question under investigation in this study is: How do students' evaluations of the quality of peer-created learning resources compare with that of domain experts? In our analysis, we also control for performance of students and the quality of the resources. Accordingly, we first investigate if, in general, student ratings to the quality of learning resources correlate with the ratings provided by domain experts. We then

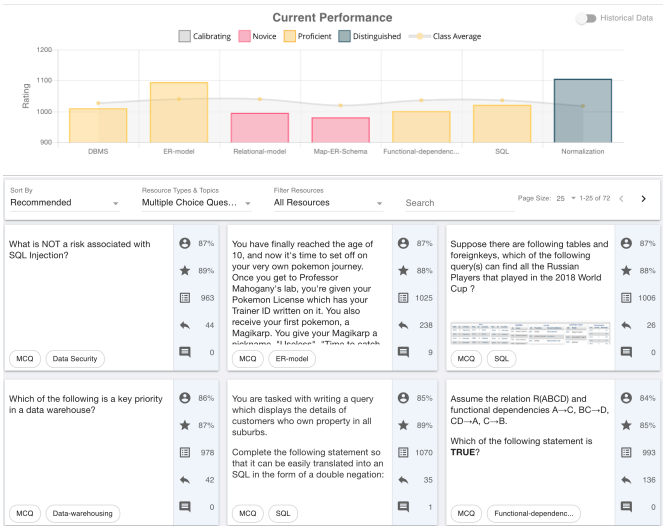

Fig. 1. Overview of the practice page in RiPPLE.

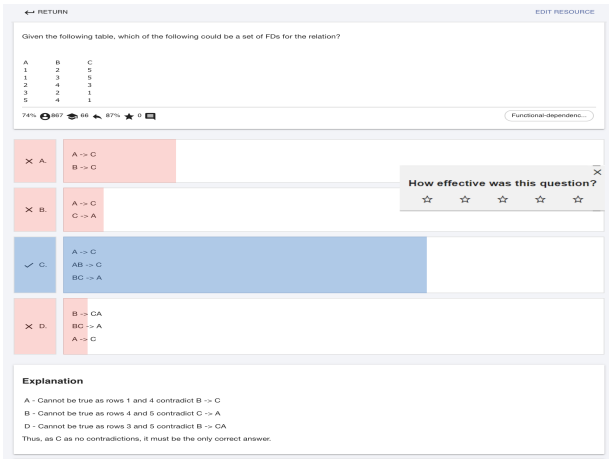

Fig. 2. Overview of the interface used for evaluating resources.

investigate if there are differences in students' ability based on their performance in the course to judge the quality of learning resources of different quality.

Tool: This study uses a course-level, discipline-agnostic platform called RiPPLE [22]. At its core, RiPPLE is an adaptive educational system that dynamically adjusts the level or type of instruction based on an individual student's ability or preferences to provide a customized learning experience [58]. Fig. 1 shows one of the main pages in RiPPLE. The upper part contains an interactive visualization widget allowing students to view an abstract representation of their knowledge state based on a set of topics associated with a course using an open learner model as outlined in [59]. The lower part of the RiPPLE screen displays learning resources recommended to a student based on their learning needs using the recommender system outlined in [60].

Instead of the common approach of relying on domain experts to develop the content for an adaptive system, RiPPLE partners with students and employs a learnersourcing approach to engage students in the creation of learning resources. In the current version of the platform, students can create MCQs, multi-answer questions, matching type questions, worked examples as well as open-ended learning resources referred to as notes.

Context: The data set used in this study is obtained from piloting RiPPLE in an on-campus computer science course on "Relational Database" at The University of Queensland (Approval from our Human Research Ethics Committee #2018000125 was received for conducting this evaluation on RiPPLE). The course incorporates many concepts that are commonly taught in an introductory course on relational databases, including conceptual database design using entity-relationship (ER) diagrams, relational models, functional dependencies, normalization, relational algebra, structured query language (SQL), and data warehousing. For maximizing students' practice and ensuring their regular engagement with RiPPLE, the course used two rubrics for computing students' final grade. In the first rubric, the final exam and RiPPLE contributed to 40% and 10% of the final grade, respectively. In the second rubric, the final exam and RiPPLE contributed to 50% and 0% of the final grade, respectively. The maximum grade obtained from these two assessment rubrics was considered as a student's final grade. The grade associated with RiPPLE was based on students' engagement with four rounds of creating and answering MCQs related to the concepts of the course at 3-weeks interval (only learning resources of type MCQ were used in the course used in this study). Participation in each round was associated with a maximum of 2 marks given that students correctly answered 10 MCQs (one mark) and authored at least one high-quality MCQ (one mark). The quality of an MCQ was determined by the voluntary ratings provided by their peers that had answered the question. The evaluation was accomplished by attributing a score to the learning resource, indicated through a number of stars out of a maximum of five. Fig. 2 presents the interface used for evaluating the quality of resources. In their rating, students were instructed to consider the following criteria: (1) The question reinforces learning from the content covered in the course; (2) The author has provided a good solution with an explanation that would be helpful to someone who answers their question incorrectly; and (3) Other options must seem plausible.

Data set: During the 13 weeks that the course was running, 521 students enrolled in this course created 2,355 MCQs, made 87,437 attempts, and provided 31,143 ratings on 2,355 peer-created learning items, which were available in the platform repository for this course. A small subset of the 2,355 created MCQs was selected to be evaluated by the domain experts and consequently to be used in the study. We took the following steps and measures to ensure that the selected questions sufficiently incorporate information on active students with different levels of performance and questions with various levels of quality. (1) Students who had answered less than 25 MCQs were considered inactive and excluded from the study, leaving 384 students for further analyses. (2) The remaining students were then divided based on their final score in the course into three groups. In accordance with Item Analysis in differentiating students [61], the highest-scoring 27% of

TABLE I
TOTAL NUMBER OF RATINGS PROVIDED BY EACH PERFORMANCE-GROUP OF STUDENTS TO EACH BIN OF QUESTIONS

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

TABLE II
AVERAGE AND STANDARD DEVIATIONS OF RATINGS TO THE THREE BINS OF QUESTIONS BY EACH GROUP OF STUDENTS AND DOMAIN EXPERTS

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

students were considered as high-performing (103 students with the mean course grade 90.5 ± 4.4), the lowest scoring 27% of students as low-performing (103 students with the mean course grade 54.8 ± 8.5), and the remaining 46% as the average-performing (178 students with the mean course grade 75.4 ± 5). (3) From the 2,355 MCQs available in the RiPPLE repository, questions that had received less than five ratings from each of the three groups of students (high-performing, average-performing, and low-performing) were excluded leaving 1,632 questions for further analysis. (4) The remaining questions were then sorted by their average ratings provided by the students in the ascending order and were divided into three groups, where each group received one-third of the questions. This led to 544 questions in each bin where low-quality questions had a mean rating of 2.85 ± 0.44, average-quality questions had a mean rating of 3.52 ± 0.13, and high-quality questions had a mean rating of 4.1 ± 0.3. From each of these three bins, 14 questions were randomly sampled to be included in the study. This formed a total of 42 questions, which were used in this study. Table I provides a summary of the number of ratings provided by each performance-group of students to each bin of questions selected for this study.

Six individuals, as the teaching staff of the course, were recruited as the domain experts to independently review the 42 questions available for this study. The team included one individual with 11 years of experience with this course, one individual with five years of experience with this course, three individuals with two years of experience with the course, and finally one individual with one year of experience with the course. The domain experts were asked to adopt the same three criteria, which were previously defined, to rate the questions. The result of the inter-rater agreement among domain experts using the intraclass correlation coefficient (ICC) [62] suggests an excellent agreement among them for evaluating the quality of learning resources (ICC[3,k] = 0.84]). For the remainder of this paper, we use the mean rating provided by instructors to a resource as the ground truth for the quality of that resource. Table II reports the averages and standard deviations of ratings by each group of students, the entire class, and the domain experts on each of the three bins of resources.

Data analysis: We conduct a regression analysis to examine the relationship between each group of student ratings and domain experts ratings. To do so, we considered the domain experts ratings as the dependent variable and the ratings given by the students as the independent variable. We report the r-value and p-value of the regressed model where r-value is the Pearson's r correlation coefficient, and the p-value is the two-sided p-value obtained from a Wald test for a hypothesis test for which the null hypothesis is that the slope of the regressed line is zero. The bigger values of Pearson's r correlation indicate a stronger correlation between the two variables. We complement the regression analysis with an error analysis where we consider the domain experts ratings as the gold standard and compute the error of the ratings provided by students (each group of students) to the quality of learning resources in each bin. In this analysis, Root Mean Squared Error (RMSE) is used to compute the error of ratings provided by students and is computed as:

$$RMSE = \sqrt{\frac{\sum_{(i)}(e_i - s_i)^2}{N}}, \quad (1)$$

where e_i and s_i are the ratings for learning resource i expressed by the domain experts and students, respectively, and N is the number of all data points in the data set for which RMSE is being reported. The smaller values of RMSE indicate the higher accuracy of ratings. To investigate the significance of the results obtained, we use bootstrapping technique to estimate a 95% confidence interval (CI) for the differences between Pearson correlations or RMSEs.

B. Results

In this section, the results of the analyses described in Section III-A are presented.

Regression analysis: Fig. 3 illustrates the relationship between the domain experts ratings and student ratings obtained from the regression analysis. In this figure, each data point represents an individual item and the regressed line demonstrates the best-fitted line obtained from the regression analysis. At the class-level (all students), the result of the Pearson's r correlation coefficient obtained from this analysis indicates a strong positive relationship between the domain experts ratings and the ratings provided by students ($r(40) = 0.78, p < .01$). At the students performance-group level, for the high-performing students, we can observe a very strong positive correlation with the domain experts ratings ($r(40) = .828$, $p < .01$). An almost similar pattern is observable for the average-performing students, but the main difference is that for this group of students, the data points are more distanced from the regressed line ($r(40) = .694$, $p < .01$), but the difference between the correlation coefficient of the high-performing and the average-performing was not significant with the 95% CI [-0.01, 0.29]. For the low-performing students, the same as the other two groups, there is a strong correlation between the ratings provided by students and the domain experts ratings ($r(40) = 0.499$, $p < 0.01$), but the difference between the correlation coefficient of the high-performing and the low-performing was significant with the 95% CI [0.17, 0.51].

Error analysis: Table III reports the RMSE of the ratings provided by the students (at the class-level and performance-group level) on each of the three bins of resources. The results

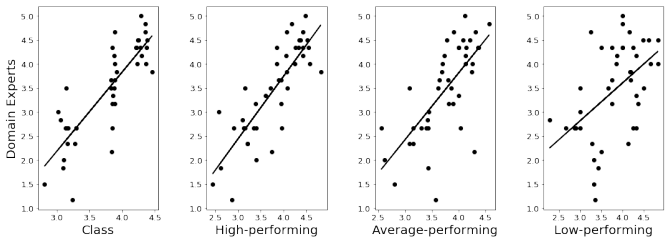

Fig. 3. Comparison of the ratings provided by the students and domain experts using the regression analysis. Here, each data points represents an individual item and the regressed line demonstrates the best-fitted line obtained from the regression analysis.

TABLE III
RMSE OF THE RATINGS BY EACH GROUP OF STUDENTS WITH REGARDS TO QUESTION QUALITY

	Low-performing	Average-performing	High-performing	Class
High-quality	0.501	0.390	0.40	0.338
Average-quality	0.916	0.801	0.596	0.649
Low-quality	1.202	1.08	0.886	0.932
All resources	0.919	0.777	0.659	0.706

indicate that the error rate across different bins of resources varies considerably; On the high-quality bin of resources, the error of the ratings provided by the high-performing and the average-performing students are almost similar and smaller than the error of the ratings provided by the low-performing students by 0.1, but the difference in RMSE was not significant between any of the three groups. On the other hand, for the average-quality and low-quality bins of resources, the high-performing students performed better in judging the quality of learning resources followed by the average-performing students, and the low-performing students. In particular, the difference between RMSE of the high-performing and the low-performing students was significant on the average-quality resources and the low-quality resources with 95% CI [0.04, 0.58] and [0.06, 0.59], respectively.

C. Summary

In summary, the findings of this study are as follows:

- Regardless of the performance level, there is a strong positive correlation between the ratings provided by the students and the domain experts ratings.
- The difference in judgmental ability was evident between different student groups. In particular, the high-performing students showed a better ability in judging the quality of learning resources followed by the average-performing students.
- Students in all three groups showed a better ability in judging the quality of the high-quality resources in comparison to the average-quality or the low-quality resources.

IV. STUDY 2: A LEARNERSOURCING-CONSENSUS APPROACH

The findings from Study 1 suggested that ratings from students may be considered an important source of data for separating high-quality resources from low-quality ones in a repository of learnersourced learning resources. However, not all students are equal in their ability to judge the quality of resources and the accuracy of ratings from different groups of students varies considerably. This makes the use of simple aggregation methods, such as averaging, unsuitable for estimating the quality of resources, as these models treat all ratings from all students equally without differentiating between their evaluation ability. Therefore, in this section, we investigate how a more advanced consensus approach that employs machine learning can be used for aggregating ratings provided by students. We also examine whether the presence of auxiliary data that are relatively accessible with low cost can be used towards improving the accuracy of the results. As such, in addition to exploring the accuracy of the consensus algorithm solely based on the presence of ratings from students, we examine whether access to information about students' academic performance can be used towards the more accurate aggregation of student ratings. In our study, students' marks on the final exam was used to approximate their performance. In practice, other options, such as students' cumulative GPA or marks from assignments during the semester, may also be considered. We also examine the accuracy of the proposed consensus algorithm as an expert-in-the-loop approach by incorporating information from domain experts on a fraction of learning resources.

TABLE IV
THE FOUR CONDITIONS FOR POPULATING THE RATING MATRIX USED IN MF

		Experts	
		Not-used	Used
Performance	Not-used	S	S + E
	Used	S + P	S + P + E

We employ MF recommendation technique as the consensus algorithm to estimate the quality of learning resources, as MF (1) works well with sparse data sets, (2) accommodates both supervised and non supervised settings, and (3) can be used to incorporate auxiliary data, as discussed in Section II-B. We use the three data sources from study 1 as explained in Section III-A to populate MF under four approaches presented in Table IV. In this table, S, P, and E refer to learnersourced student ratings data, student performance data, and domain experts ratings data, respectively. 'Used' ('Not-used') indicates whether the supplementary data was being used (Not being used) by MF. In addition, using $'+'$ indicates the simultaneous use of data sources. For example, $S + E$ means that student ratings and domain experts ratings are used together by the algorithm. In what follows, we first represent more details on MF. We then discuss the ways that it is employed under each of the four approaches presented in Table IV.

A. Method

*1) **MF**:* In this section, more explicit details are provided on the MF recommendation algorithm for aggregating student ratings and inferring the right quality of learning resources. In this context, identifying the quality of learning resources from

	Student ID	Resource ID	Quality Rating
Train	u_1	q_1	1
	u_3	q_1	2
	u_4	q_1	1
	u_1	q_2	3
	u_2	q_2	3
	u_3	q_2	4
	u_1	q_3	5
	u_2	q_3	4
	u_4	q_3	3
Test	u_1	q_4	2
	u_3	q_4	3
	u_1	q_5	4
	u_2	q_5	4
	u_4	q_5	3

Students ratings (s)

Student ID	Score	Performance group
u_1	7	High-performing
u_2	3	Low-performing
u_3	5	Average-performing
u_4	2	Low-performing

Student performance data (P)

	Resource ID	Quality Rating
Train	q_1	2
	q_3	5
	q_4	?
Test	q_5	?

Domain expert ratings (E)

Fig. 4. An illustrative example of three data sources with 4 students, 5 learning resources and 14 ratings given by students.

learnersourced ratings can be considered as a rating prediction problem, where a student, learning resource and quality rating could be treated as a user, item, and rating, respectively. In what follows, let $U_N = \{u_1 \dots u_N\}$ denote a set of students who are enrolled in a course, where u_i refers to an arbitrary student. Let $Q_M = \{q_1 \dots q_M\}$ present the repository of learning resources that are available to students in a course, where q_j refers to an arbitrary learning resource. In addition, a two-dimensional array $R_{N\times M}$ provides information on students' perceptions of the quality of learning resources; whenever a student u_i expresses a rating for the quality of learning resource q_j, the rating is stored in r_{ij}. Let $H_{N\times K}$ represent the latent factors underlying students behavior, where h_i is a vector of latent factors representing student u_i. Similarly, let $L_{M\times K}$ represent the latent factors of a learning resource set, where l_j is a vector of latent factors representing item q_j. After the mapping of students and learning resources to the latent factors, the rating of a student u_i for a learning resource q_j can be approximated as:

$$\hat{r}_{ij} = l_j^{\mathsf{T}} h_i = \sum_{k=1}^{K} l_{jk} h_{ik}. \tag{2}$$

Matrix $\hat{R} = \{\hat{r}_{ij}\}$ is then used to capture all predicted ratings that students give to a set of resources, with the elements given by Equation 2. The goal of MF is to learn the matrices H and L, which are used to compute values for $\hat{R}$. To learn H and L matrices, MF minimizes the following

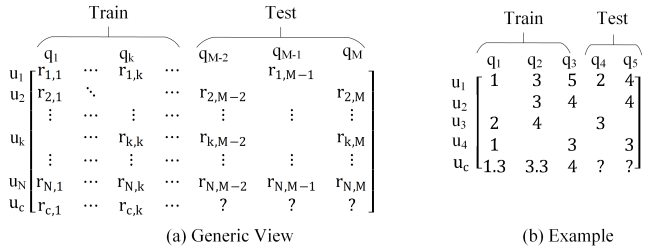

Fig. 5. Overview of the rating matrix passed to MF using learnersourced student ratings (S).

regularized squared error term on the set of known ratings:

$$\sum_{(i,j)\in R_{train}} (r_{ij} - l_j^{\mathsf{T}} h_i)^2 + \lambda({l_j}^2 + {h_i}^2), \tag{3}$$

where $(u_i, q_j) \in R_{train}$ represents (u_i, q_j) pairs such that the rating of student u_i for item q_j is present in the training data set and λ is a parameter controlling the extent of the regularization.

*2) **Using MF for rating aggregation**:* In what follows, the four approaches presented in Table IV are explained. Fig. 4 presents an illustrative example of three data sources based on four students, five questions, and fourteen ratings, which have been split to train and test sets. Based on the reported scores in the student performance data (P), u_1 is a high-performing student, u_3 is an average-performing student, and u_2 and u_4 are low-performing students.

Approach 1: learnersourced student ratings (S). This approach introduces a dummy student (u_c) based on the average of ratings from the entire class. Fig. 5(a) illustrates a generic view of the rating matrix generated by this approach. In this matrix, the first N rows are populated by the ratings provided by individual students to the quality of the resources, where $r_{i,j}$ is the rating given by an arbitrary student u_i to an arbitrary learning resource q_j. The $(N+1)^{th}$ row is dedicated to representing ratings from u_c, where $r_{c,j}$ is computed as the average of all given ratings by students to q_j. MF is then employed to approximate how u_c would rate the questions that are in the test set. This approach can be considered as an unsupervised approach, where the inferred ratings are computed in the absence of information about the true ratings of the resources. Fig. 5(b) illustrates the MF rating matrix generated based on the provided example. The ratings given to the three resources in the train set (q_1, q_2, q_3) are used to compute the auxiliary ratings given by u_c. As an example, $r_{c,1} = \frac{1+2+1}{3} = 1.3$. MF uses this information to infer u_c's ratings for q_4 and q_5.

Approach 2: learnersourced student ratings and student performance data (S+P). This approach introduces a dummy student (u_h) based on the average of the ratings from high-performing students. Fig. 6(a) illustrates a generic view of the rating matrix generated by this approach. The ratings given by individual students populate the first N rows of R. The $(N+1)^{th}$ row is dedicated to representing ratings from u_h. For instance, $r_{h,j}$ is computed as the average of the ratings given to q_j by high-performing students. MF is then employed

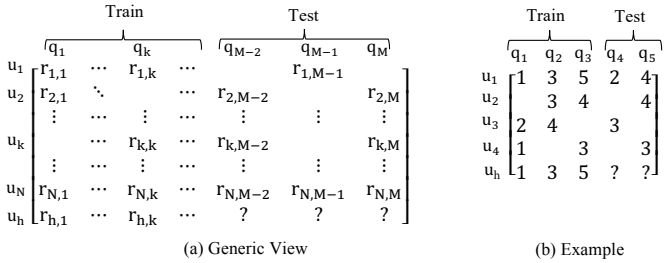

Fig. 6. Overview of the rating matrix passed to MF when simultaneously using student ratings data and performance data ($S + P$).

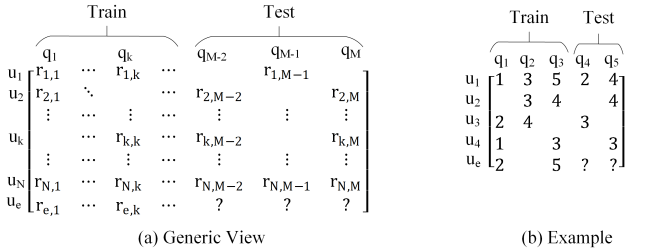

Fig. 7. Overview of the rating matrix passed to MF when simultaneously using student ratings data and the domain experts ratings data ($S + E$).

to approximate how u_h would rate the questions that are in the test set. This model can also be considered as an unsupervised machine learning approach. Fig. 6(b) illustrates the MF rating matrix generated based on the provided example.

Approach 3: learnersourced student ratings and domain experts ratings (S+E). This approach introduces a dummy expert (u_e) based on the average of ratings provided by domain experts. Fig. 7(a) illustrates a generic view of the rating matrix generated by this approach. The first N rows are populated by the ratings from individual students to the quality of learning resources. The $(N + 1)^{th}$ row is dedicated to representing ratings from u_e, where $r_{e,h}$ is the average of ratings given by domain experts to the quality of q_j. MF is then employed to approximate how u_e would rate the questions that are in the test set. This approach can be considered as a semi-supervised approach, where the inferred ratings are computed based on the information from the true ratings of a fraction of the resources. Fig. 7(b) illustrates the MF rating matrix generated based on the provided example. The ratings given to the resources by the students and the domain experts in the train set are used by MF to infer u_e's rating for q_4 and q_5 in the test set.

Approach 4: learnersourced student ratings, student performance data, and domain experts ratings (S+P+E). This approach introduces one dummy student (u_h) based on the average of ratings from high-performing students. It also introduces one dummy expert (u_e) based on ratings from domain experts. These dummy participants are equivalent to those introduced in the second and third approaches. Fig. 8(a) represents an overview of the rating matrix generated by this approach. The ratings from individual students populate the first N rows. The $(N+1)^{th}$ is dedicated to representing ratings from u_h, as explained in approach 2. The $(N + 2)^{th}$ row of

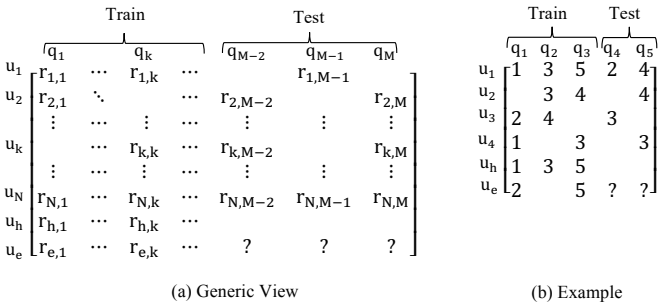

Fig. 8. Overview of the rating matrix passed to MF when using student ratings data, performance data and the domain experts ratings data ($S + P + E$).

R is dedicated to representing ratings from u_e as explained in approach 3. MF is then employed to approximate how u_e would rate the questions that are in the test set. This approach can also be considered as a semi-supervised approach, where the inferred ratings are computed based on the information from the true ratings of a fraction of the resources. Fig. 8(b) illustrates the MF rating matrix in this approach based on the provided example. The ratings given to the three resources in the train set (q_1, q_2, q_3), as well as student performance data are used to compute the auxiliary ratings given by u_h. Also, the ratings given by the domain experts to q_1 and q_3 are added to the last row of the rating matrix. MF then uses this information to infer u_e's rating for q_4 and q_5.

B. Evaluation

To evaluate the proposed rating aggregation model, we use the three data sources, from study 1 described in Section III-A, based on the 42 selected learning resources and conducted two analyses: Baseline comparison and unsupervised vs. semi-supervised approaches comparison. In all analyses, we use an extension of MF called Biased-MF proposed by Koren [63] that incorporates mean normalization and a bias parameter for each student and learning resource in ratings. We use RMSE as our evaluation metric, where it measures the differences between the predicted ratings by MF and the domain experts ratings as the actual data. We use the MF implementation of MyMediaLite [64] for all of our conducted studies. The reported RMSE is the result of using five-fold cross-validation on the data sets.

1) Baseline comparison: This analysis aims to investigate whether in the absence of ratings from domain experts, using unsupervised MF-based consensus approaches can be used to improve the accuracy of approximating the quality of learning resources compared to just relying on human intelligence data and using averaging aggregation. For the analysis, we first compare the accuracy of the first MF-based approach, ('S'), with the results obtained from applying averaging aggregation on the student ratings data reported in table III (study 1). We then investigate whether incorporating supplementary information about student performance based on the second MF-based approach, ('S+P'), impacts the accuracy of the approximated ratings. Table V summarizes the results for the accuracy of the two unsupervised MF-based approaches, ('S' and 'S+P'),

TABLE V
RMSE VALUES FOR THE TWO UNSUPERVISED MF-BASED APPROACHES ('S' AND 'S+P') AND THE CORRESPONDING RESULT FROM STUDY 1

Participants	Mean (Study 1)	MF (S) (Study 2)	MF (S+P) (Study 2)
RMSE	0.699	0.657	0.595

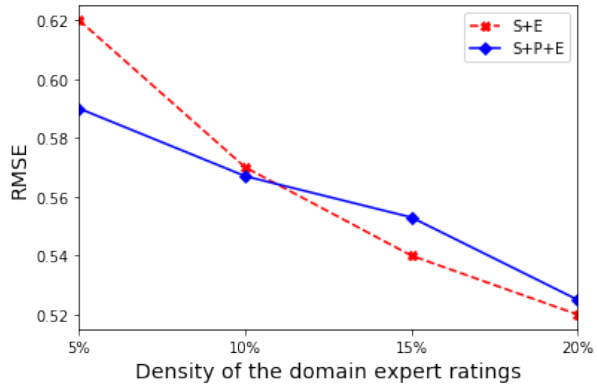

Fig. 9. Comparing the accuracy of the two semi-supervised MF-based approaches with varying levels of density.

in determining the quality of learning resources. In order to facilitate the comparisons, the corresponding result from study 1 is also added to table V.

The reported results illustrate that applying the first MF-based approach, ('S'), on the student ratings data leads to higher accuracy ($RMSE = 0.657$) compared to applying averaging aggregation on the ratings from students in the class ($RMSE = 0.699$). This RMSE is very close to the RMSE obtained from applying averaging aggregation on the ratings given by high-performing students suggesting that without having any additional data about students' performance, by combining machine learning and human intelligence, we can obtain the highest possible accuracy that is attainable by using averaging aggregation on the ratings provided by students. With regards to the second MF-based approach, ('S+P'), the reported RMSE ($RMSE = 0.595$) indicates that, incorporating student performance data in the base work of MF, not only leads to a higher accuracy that can be obtained by applying averaging aggregation on the high-performing student ratings ($RMSE = 0.659$), but also provides an accuracy beyond what is achievable by just relying on the ratings provided students using the first MF-based approach ('S').

2) Unsupervised vs. semi-supervised approaches: This analysis aims to investigate whether ratings from domain experts on a limited number of learning resources based on the two semi-supervised approaches ('S+E' and 'S+P+E') lead to higher accuracy in approximating the quality of learning resources. For this comparison, the density of the domain experts ratings in the semi-supervised approaches is varied from 5% to 20% with the step value of 5%. Fig. 9 represents the RMSE values obtained from employing the two semi-supervised approaches.

Fig. 9 indicates that by having the domain experts ratings on 5% of learning resources, 'S+E' attains higher accuracy ($RMSE = 0.620$) compared to having no ratings from the domain experts ($RMSE = 0.657$); On the other hand, having this amount of ratings from the domain experts only slightly improves the RMSE attained by 'S+P+E' approach ($RMSE = 0.590$) compared to having no ratings from the domain experts ($RMSE = 0.595$). By changing the density of the domain experts ratings from 5% to 10%, the RMSE value attained by both 'S+E' and 'S+P+E' approaches decreases sharply from 0.62 to 0.57 for 'S+E' approach and from 0.590 to 0.565 for 'S+P+E' approach. This density level is likely to be highly data set dependent; In the case of our data set, 5% density level corresponds to the ratings from the domain experts to only two learning resources and this amount of information from the domain experts did not allow MF to appropriately calibrate students' contribution based on their similarity to the domain experts ratings. We speculate that in bigger data sets, this turning point in accuracy can occur at a lower density level.

On the other hand, comparing 'S+E' and 'S+P+E' suggests that, in both of these approaches, an increase in the density of the domain experts ratings improves RMSE; however, for 10% density, 'S+E' performs similar to 'S+P+E' and beyond 10%, 'S+E' slightly outperforms 'S+P+E'. This finding suggests that, in the absence of ratings from domain experts, student performance data works as a proxy for the reliability of students; however, as soon as some ratings from the domain experts are added to the base work of MF, it no longer requires a proxy for reliability, and instead, MF can determine the reliability of individual students based on the similarity of their ratings with the ratings given by the domain experts.

C. Summary

In summary, the findings of this study are as follows:

- Using unsupervised learnersourcing consensus approaches based on MF can improve the accuracy of estimating the quality of learning resources compared to using simple statistical aggregations.
- Using MF that leverages student performance data can lead to higher accuracy compared to using MF solely with the ratings provided by students based on the unsupervised approaches.
- Using semi-supervised consensus approaches based on MF that incorporates ratings from the domain experts on a limited number of learning resources can considerably improve the accuracy in approximating the quality of learning resources compared to the unsupervised approaches.

V. DISCUSSION AND CONCLUSIONS

The overarching goal of this paper is to contribute to the understanding of whether and how learnersourcing can be used for evaluating the quality of learning resources. To do so, we conducted two studies using data sets obtained from a learnersourcing online learning platform called RiPPLE that relies on students for the development of learning resources as well as evaluating the quality of learning resources that exist in the repository of the system. In particular, the first study

(Study 1) investigated the plausibility of involving students in evaluating the quality of learning resources. The results of Study 1 provided evidence that there is a strong positive correlation between the ratings provided by students and the domain experts ratings and that the competency of students in judging the quality of learning resources was increased with the students' performance in the course. Particularly, there were evident differences between high-performing students and low-performing students across all conducted experiments in judging the quality of learning resources. This is not surprising, as by definition, we expect the high-performing students to perform better compared to other students. Motivated by the findings from study 1, which suggested differences among students in their ability to judge the quality of learning resources, and extensive work within the community of crowdsourcing for optimal integration of crowdsourced data using machine learning, in Study 2, we proposed a consensus approach based on matrix factorization (MF) for aggregating student ratings. Results suggest that using our proposed consensus approaches, instead of the simple statistical aggregations, can improve the accuracy of determining the quality of learning resources. In addition, using auxiliary data in the form of student performance or limited ratings from domain experts, which are relatively accessible with low cost, can further improve the results.

There are several limitations in the current work. One important limitation is that we conducted our evaluations in the context of one course with one type of resource on a limited number of peer-created resources. This limits the generalizability of our findings to other kinds of disciplinary learning resources that are common in other learning management systems, such as open-ended questions, exemplars, or tutorial videos. As a first step, future work includes replicating the conducted studies with courses across disciplines to investigate the generalizability of our current findings. The second limitation of the current study is that it used a single-scale evaluation metric for capturing students' evaluation of the quality of resources across multiple dimensions. Future works aim to investigate the impact of using multi-criteria rubrics for capturing students' evaluations of the quality of a resource. Finally, future work also includes investigating how digital footprints of students in the platform, such as time taken to evaluate a resource, may be leveraged to further improve the results.

Implications: While there are many benefits associated with learnersourcing, there are also concerns associated with sharing repositories of learnersourced resources with students, as some of the resources may be poorly worded or incorrect. A potential solution for addressing this challenges is to develop a formal evaluation process that partners with students as decision-makers in deciding whether or not a resource authored by a student should be added to the resource repository of the system. However, in practice, accommodating a formal evaluation process in learnersourcing platforms introduces a number of complex and interdependent requirements which has not been considered by this paper and needs to be addressed by future research. Some of these requirements are listed in the following.

Explainable consensus algorithms. Most of the state-of-the-art consensus approaches developed within crowdsourcing community rely on black-box machine learning algorithms (e.g., [37], [65]) for aggregating crowdsourced decisions. Using these machine learning algorithms have considerably improved the accuracy of decision aggregation compared with simple statistical models, such as averaging aggregation. However, generally, these machine learning-based consensus approaches are not understandable and transparent in terms of how decisions from each individual are weighted and how the final decision was made [66]. While some studies suggest that explainable AI (XIA) [67] is not always required [68], the use of black-box outcomes seems to be inadequate for educational settings where educators strive to enable students to develop their own vision, reasoning, and appreciation for inquiry and investigation, and fairness. Much of the existing works on the need for open and XIA models in education has focused on open learner models [69], educational recommender systems [70], and learning analytics dashboards [71]. An interesting direction that can be followed in the future would be (1) investigating the design of consensus algorithms that are accurate and explainable, (2) quantifying the evaluation ability of students in an accurate and transparent way, and (3) extending open learner models for communicating students' reliability in evaluating peer-created learning resources.

Feedback. Another essential aspect of engaging students in a formal evaluation process is the ability to provide training, support, and feedback mechanisms to help students develop their evaluative judgment capacity at scale as well as providing them with a more personalized and tailored learning experience. Interesting future directions that can be followed with this regard include (1) investigating appropriate methods for providing personalized training and feedback to students based on their learnersourced contributions and interaction behaviors to support learning and development of evaluative judgment, (2) developing behavior-based recommendation and consensus algorithms for generating personalized results that meet the needs of the learner, (3) studying the advantages, if any, of the personalized consensus-based feedback.

Optimal expert involvement. In a formal evaluation process, given the limited availability of domain experts, development of expert in-the-loop consensus approaches necessitates the selection of the most informative learning resources to be evaluated by domain experts. As such, an interesting direction to be followed in the future would be to investigate how to optimally utilize the limited availability of experts based on spot-checking algorithms [72] or active learning methods [73] to enhance the reliability and accuracy of the consensus approaches.

REFERENCES

[1] J. Kim, "Learnersourcing: Improving learning with collective learner activity," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, USA, 2015.

[2] J. Kim, P. J. Guo, C. J. Cai, S.-W. Li, K. Z. Gajos, and R. C. Miller, "Data-driven interaction techniques for improving navigation of educational videos," in *Proc. 27th Annu. ACM Symp. User Interface Software Technology*, Oct. 2014, pp. 563–572, doi: 10.1145/2642918.2647389.

[3] X. Wang, S. T. Talluri, C. Rose, and K. Koedinger, "UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities," in *Proc. 6th (2019) ACM Conf. Learning@ Scale*, Jun. 2019, pp. 1–10, doi: 10.1145/3330430.3333614.

[4] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan, "Axis: Generating explanations at scale with learnersourcing and machine learning," in *Proc. 3rd (2016) ACM Conf. Learning@ Scale*, Apr. 2016, pp. 379–388, doi: 10.1145/2876034.2876042.

[5] N. T. Heffernan, K. S. Ostrow, K. Kelly, D. Selent, E. G. Van Inwegen, X. Xiong, and J. J. Williams, "The future of adaptive learning: Does the crowd hold the key?" *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 615–644, Feb. 2016, doi: 10.1007/s40593-016-0094-z.

[6] S. Doroudi, "Integrating human and machine intelligence for enhanced curriculum design," Ph.D. dissertation, Air Force Res. Lab., Dept. Comput. Sci., CMU, Pittsburgh, PA, USA, 2019.

[7] T. T. Hills, "Crowdsourcing content creation in the classroom," *J. Comput. Higher Educ.*, vol. 27, no. 1, pp. 47–67, Jan. 2015, doi: 10.1007/s12528-015-9089-2.

[8] M. Neary, "Student as producer: A pedagogy for the avant-garde?" *Learn. Exchange*, vol. 1, no. 1, Dec. 2010.

[9] N. M. Lambert and B. L. McCombs, *How students learn: Reforming schools through learner-centered education*. Washington, DC, USA: American Psychological Assoc., 1998, doi: 10.1037/10258-000.

[10] D. C. Edelson, D. N. Gordin, and R. D. Pea, "Addressing the challenges of inquiry-based learning through technology and curriculum design," *J. Learn. Sci.*, vol. 8, no. 3-4, pp. 391–450, Jul. 1999, doi: 10.1207/s15327809jls08034-3.

[11] J. Hamer, Q. Cutts, J. Jackova, A. Luxton-Reilly, R. McCartney, H. Purchase, C. Riedesel, M. Saeli, K. Sanders, and J. Sheard, "Contributing student pedagogy," *ACM SIGCSE Bull.*, vol. 40, no. 4, pp. 194–212, Nov. 2008, doi: 10.1145/1473195.1473242.

[12] H. Purchase, J. Hamer, P. Denny, and A. Luxton-Reilly, "The quality of a PeerWise MCQ repository," in *Proc. 12th Australasian Conf. Computing Education*, Jan. 2010, pp. 137–146, doi: 10.5555/1862219.1862238.

[13] B. S. Bloom, M. Englehart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of educational objectives: Handbook I, Cognitive domain*. New York, NY, USA: David McKay, 1956.

[14] M. Barak and S. Rafaeli, "On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning," *Int. J. Human-Comput. Stud.*, vol. 61, no. 1, pp. 84–103, Jul. 2004, doi: 10.1016/j.ijhcs.2003.12.005.

[15] S. W. Draper, "Catalytic assessment: Understanding how MCQs and EVS can foster deep learning," *Brit. J. Educ. Technol.*, vol. 40, no. 2, pp. 285–293, Feb. 2009, doi: 10.1111/j.1467-8535.2008.00920.x.

[16] P. Denny, J. Hamer, A. Luxton-Reilly, and H. Purchase, "PeerWise: Students sharing their multiple choice questions," in *Proc. 4th Int. Workshop Computing Education Research*, Sep. 2008, pp. 51–58, doi: 10.1145/1404520.1404526.

[17] S. P. Bates, R. K. Galloway, and K. L. McBride, "Student-generated content: Using PeerWise to enhance engagement and outcomes in introductory physics courses," in *AIP Conf. Proc.*, Aug. 2012, pp. 123–126, doi: 10.1063/1.3680009.

[18] "Constructivism: Implications for the design and delivery of instruction," in *Handbook of Research for Educational Communications and Technology*. New York, NY, USA: Simon and Schuster, 1996, pp. 170–198.

[19] J. D. Raskin, "Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism," *Amer. Commun. J.*, vol. 5, no. 3, pp. 1–25, 2002.

[20] S. Tackett, M. Raymond, R. Desai, S. A. Haist, A. Morales, S. Gaglani, and S. G. Clyman, "Crowdsourcing for assessment items to support adaptive learning," *Med. teacher*, vol. 40, no. 8, pp. 838–841, Aug. 2018, doi: 10.1080/0142159X.2018.1490704.

[21] P. Denny, A. Luxton-Reilly, and J. Hamer, "The PeerWise system of student contributed assessment questions," in *Proc. 10th Australasian Conf. Computing Education*, Jan. 2008, pp. 69–74.

[22] H. Khosravi, K. Kitto, and W. Joseph, "RiPPLE: A crowdsourced adaptive platform for recommendation of learning activities," *J. Learn. Analytics*, vol. 6, no. 3, pp. 91–105, Dec. 2019, doi: 10.18608/jla.2019.63.12.

[23] J. L. Walsh, B. H. Harris, P. Denny, and P. Smith, "Formative student-authored question bank: Perceptions, question quality and association with summative performance," *Postgraduate Med. J.*, vol. 94, no. 1108, pp. 97–103, Feb. 2018, doi: 10.1136/postgradmedj-2017-135018.

[24] P. Denny, A. Luxton-Reilly, and B. Simon, "Quality of student contributed questions using PeerWise," in *Proc. 11th Australasian Conf. Computing Education*, Jan. 2009, pp. 55–63.

[25] J. Tai, R. Ajjawi, D. Boud, P. Dawson, and E. Panadero, "Developing evaluative judgement: Enabling students to make decisions about the quality of work," *Higher Educ.*, vol. 76, no. 3, pp. 467–481, Sep. 2018, doi: 10.1007/s10734-017-0220-3.

[26] D. Boud, R. Ajjawi, P. Dawson, and J. Tai, *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work*. Evanston, IL, USA: Routledge, 2018, doi: 10.4324/9781315109251.

[27] H. Khosravi, G. Gyamfi, B. E. Hanna, and J. Lodge, "Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system," in *Proc. 10th Int. Conf. Learning Analytics and Knowledge*, Mar. 2020, pp. 83–88, doi: 10.1145/3375462.3375532.

[28] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" in *Proc. VLDB Endowment*, Aug. 2017, pp. 541–552, doi: 10.14778/3055540.3055547.

[29] S. Abdi, H. Khosravi, S. Sadiq, and G. Demartini, "Code supplement for evaluating the quality of learning resources," 2019. [Online]. Available: https://github.com/solmazabdi/Learnersourcing-twoStudies.git

[30] F.-Y. Yu, Y.-H. Liu, and T.-W. Chan, "A web-based learning system for question-posing and peer assessment," *Innov. Educ. Teaching Int.*, vol. 42, no. 4, pp. 337–348, Nov. 2005, doi: 10.1080/14703290500062557.

[31] P. Denny, B. Hanks, and B. Simon, "PeerWise: Replication study of a student-collaborative self-testing web service in a us setting," in *Proc. 41st ACM Technical Symp. Computer Science Education*, Mar. 2010, pp. 421–425, doi: 10.1145/1734263.1734407.

[32] R. Grainger, W. Dai, E. Osborne, and D. Kenwright, "Medical students create multiple-choice questions for learning in pathology education: A pilot study," *BMC Med. Educ.*, vol. 18, no. 1, p. 201, Aug. 2018, doi: 10.1186/s12909-018-1312-1.

[33] P. Denny, J. Hamer, and A. Luxton-Reilly, "Students sharing and evaluating MCQs in a large first year engineering course," in *20th Annu. Conf. Australasian Association for Engineering Education*, Dec. 2009, p. 575.

[34] A. E. Kay, J. Hardy, and R. K. Galloway, "Student use of PeerWise: A multi-institutional, multidisciplinary evaluation," *Brit. J. Educ. Technol.*, vol. 51, Feb. 2020, doi: 10.1111/bjet.12754.

[35] A. Farasat, A. Nikolaev, S. Miller, and R. Gopalsamy, "Crowdlearning: Towards collaborative problem-posing at scale," in *Proc. 4th ACM Conf. Learning@ Scale*, Apr. 2017, pp. 221–224, doi: 10.1145/3051457.3053990.

[36] E. Karataev and V. Zadorozhny, "Adaptive social learning based on crowdsourcing," *IEEE Trans. Learn. Technol.*, vol. 10, no. 2, pp. 128–139, Apr. 2016, doi: 10.1109/TLT.2016.2515097.

[37] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 155–164, doi: 10.1145/2566486.2567989.

[38] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: A word processor with a crowd inside," in *Proc. 23nd Annu. ACM Symp. User Interface Software and Technology*, Oct. 2010, pp. 313–322, doi: 10.1145/1866029.1866078.

[39] U. Gadiraju, G. Demartini, R. Kawase, and S. Dietze, "Human beyond the machine: Challenges and opportunities of microtask crowdsourcing," *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 81–85, Jul./Aug. 2015, doi: 10.1109/MIS.2015.66.

[40] J. Tu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, and M. Guo, "Multi-label crowd consensus via joint matrix factorization," *Knowl. Inf. Syst.*, vol. 62, no. 4, pp. 1341–1369, Apr. 2020, doi: 10.1007/s10115-019-01386-7.

[41] J. Wang, P. G. Ipeirotis, and F. Provost, "Managing crowdsourcing workers," in *2011 Winter Conf. Business Intelligence*, Mar. 2011, pp. 10–12.

[42] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing," in *Workshops 25th AAAI Conf. Artificial Intelligence*, Aug. 2011, pp. 43–48.

[43] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: Using implicit behavioral measures to predict task performance," in *Proc. 24th Annu. ACM Symp. User Interface Software Technology*, Oct. 2011, pp. 13–22, doi: 10.1145/2047196.2047199.

[44] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. 22nd Int. Conf Advances Neural Information Processing Systems*, Dec. 2009, pp. 2035–2043.

[45] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, no. 43, pp. 1297–1322, Aug. 2010.

[46] W. Bi, L. Wang, J. T. Kwok, and Z. Tu, "Learning to predict from crowdsourced data," in *Proc. 30th Conf. Uncertainty in Artificial Intelligence*, Jul. 2014, pp. 82–91.

[47] A. Kurve, D. J. Miller, and G. Kesidis, "Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 794–809, Mar. 2014, doi: 10.1109/TKDE.2014.2327026.

[48] J. Hernández-González, I. Inza, and J. A. Lozano, "A note on the behavior of majority voting in multi-class domains with biased annotators," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 1, pp. 195–200, Jan. 2018, doi: 10.1109/TKDE.2018.2845400.

[49] H. J. Jung and M. Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Workshops 26th AAAI Conf. Artificial Intelligence*, Jul. 2012, pp. 101–106.

[50] H. Morise, S. Oyama, and M. Kurihara, "Collaborative filtering and rating aggregation based on multicriteria rating," in *2017 IEEE Int. Conf. Big Data*, Dec. 2017, pp. 4417–4422, doi: 10.1109/BigData.2017.8258477.

[51] H. Morise, S. Oyama, and M. Kurihara, "Bayesian probabilistic tensor factorization for recommendation and rating aggregation with multicriteria evaluation data," *Expert Syst. Appl.*, vol. 131, pp. 1–8, Oct. 2019, doi: 10.1016/j.eswa.2019.04.044.

[52] J. R. Wright, C. Thornton, and K. Leyton-Brown, "Mechanical TA: Partially automated high-stakes peer grading," in *Proc. 46th ACM Technical Symp. Computer Science Education*, Mar 2015, pp. 96–101, doi: 10.1145/2676723.2677278.

[53] V. Shnayder and D. C. Parkes, "Practical peer prediction for peer assessment," in *4th AAAI Conf. Human Computation and Crowdsourcing*, Oct. 2016, pp. 199–208.

[54] D. K. Wind, R. M. Jørgensen, and S. L. Hansen, "Peer feedback with Peergrade," in *ICEL 2018 13th Int. Conf. e-Learning*, Jul. 2018, p. 184.

[55] H. Purchase and J. Hamer, "Peer-review in practice: Eight years of Aropä," *Assessment & Eval. in Higher Educ.*, vol. 43, no. 7, pp. 1146–1165, Oct. 2018, doi: 10.1080/02602938.2018.1435776.

[56] D. E. Paré and S. Joordens, "Peering into large lectures: Examining peer and expert mark agreement using peerScholar, an online peer assessment tool," *J. Comput. Assisted Learn.*, vol. 24, no. 6, pp. 526–540, Dec. 2008, doi: 10.1111/j.1365-2729.2008.00290.x.

[57] L. Chen, Z. Wu, J. Cao, G. Zhu, and Y. Ge, "Travel recommendation via fusing multi-auxiliary information into matrix factorization," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 1–24, Jan. 2020, doi: 10.1145/3372118.

[58] H. Khosravi, S. Sadiq, and D. Gašević, "Development and adoption of an adaptive learning system: Reflections and lessons learned," in *Proc. 51st ACM Technical Symp. Computer Science Education*, Mar. 2020, pp. 58–64, doi: 10.1145/3328778.3366900.

[59] S. Abdi, H. Khosravi, S. Sadiq, and D. Gašević, "A multivariate elo-based learner model for adaptive educational systems," in *Proc. 12th Int. Conf. Educ. Data Mining*, Jul. 2019, pp. 462–467.

[60] H. Khosravi, K. Kitto, and K. Cooper, "RiPLE: Recommendation in peer-learning environments based on knowledge gaps and interests," *J. Educ. Data Mining*, vol. 9, no. 1, pp. 42–67, Sep. 2017.

[61] S. Matlock-Hetzel, "Basic concepts in item and test analysis," presented at the Annu. Meeting Southwest Educational Research Association, Austin, TX, USA, Jan. 23–25, 1997.

[62] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bull.*, vol. 86, no. 2, p. 420, Mar. 1979, doi: 10.1037/0033-2909.86.2.420.

[63] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, Aug. 2008, pp. 426–434, doi: 10.1145/1401890.1401944.

[64] Z. Gantner, S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "MyMediaLite: A free recommender system library," in *Proc. 5th ACM Conf. Recommender Systems*, Oct. 2011, pp. 305–308, doi: 10.1145/2043932.2043989.

[65] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 469–478, doi: 10.1145/2187836.2187900.

[66] A. Darvishi, H. Khosravi, and S. Sadiq, "Utilising learnersourcing to inform design loop adaptivity," in *European Conf. Technology Enhanced Learning*, Mar. 2020, pp. 332–346, doi: 10.1007/978-3-030-57717-9_24.

[67] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proc. 2019 CHI Conf. Human Factors Computer Systems*, Jul. 2019, pp. 1–15, doi: 10.1145/3290605.3300831.

[68] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important? a study of deployed, low-cost intelligent interactive systems," in *Proc. 2012 ACM International Conf. Intelligent User Interfaces*, Feb. 2012, pp. 169–178, doi: 10.1145/2166966.2166996.

[69] S. Bull and J. Kay, "Open learner models," in *Advances in Intelligent Tutoring Systems*. Berlin, Germany: Springer, 2010, pp. 301–322, doi: 10.1007/978-3-642-14363-21_5.

[70] S. Abdi, H. Khosravi, S. Sadiq, and D. Gašević, "Complementing educational recommender systems with open learner models," in *Proc. 10th Int. Conf. Learning Analytics Knowledge*, Sep. 2020, pp. 360–365, doi: 10.1145/3375462.3375520.

[71] S. Shabaninejad, H. Khosravi, M. Indulska, A. Bakharia, and P. Isaias, "Automated insightful drill-down recommendations for learning analytics dashboards," in *Proc. 10th Int. Conf. Learning Analytics and Knowledge*, Mar. 2020, pp. 41–46, doi: 10.1145/3375462.3375539.

[72] W. Wang, B. An, and Y. Jiang, "Optimal spot-checking for improving evaluation accuracy of peer grading systems," in *Proc. 32th AAAI Conf. Artificial Intelligence*, Feb. 2018, pp. 833–840.

[73] T.-Y. Yang, R. S. Baker, C. Studer, N. Heffernan, and A. S. Lan, "Active learning for student affect detection," in *Proc. 12th Int. Conf. Educ. Data Mining*, Jul. 2019, pp. 208–217.

Solmaz Abdi is currently working towards her Ph.D. degree in the school of Information Technology and Electrical Engineering and Institute for Teaching and Learning Innovation at the University of Queensland. She received her M.Sc degree in electronic engineering from Amirkabir University of Technology, Tehran, Iran. Her research interests include machine learning, educational data mining, and recommender systems for technology enhanced learning.

Hassan Khosravi is a Senior Lecturer in the Institute for Teaching and Learning Innovation and an Affiliate Academic in the School of Information Technology and Electrical Engineering at The University of Queensland. Hassan holds a Ph.D. in Computer Science from Simon Fraser University in Canada and a Masters in Computer Science from Amirkabir University of Technology in Iran. In his research, he draws on theoretical insights from the learning sciences and exemplary techniques from the fields of human-computer interaction, learning analytics and crowdsourcing to design, implement, validate and deliver sociotechnical solutions that contribute to the delivery of learner-centred, data-driven learning at scale.

Shazia Sadiq is a Professor in the School of Information Technology and Electrical Engineering at The University of Queensland, Brisbane, Australia. She is part of the Data and Knowledge Engineering (DKE) research group and is involved in teaching and research in databases and information systems. Shazia holds a Ph.D. from The University of Queensland in Information Systems and a Masters degree in Computer Science from the Asian Institute of Technology, Bangkok, Thailand. Her main research interests are innovative solutions for Business Information Systems that span several areas including business process management, governance, risk and compliance, data quality management, workflow systems, and service science.

Gianluca Demartini is an Associate Professor in the School of Information Technology and Electrical Engineering at The University of Queensland, Brisbane, Australia. Gianluca holds a Ph.D. in Computer Science from Leibniz University of Hanover, Germany. He is part of the Data Science research group and his research has been supported by the Australian Research Council, the UK Engineering and Physical Sciences Research Council (EPSRC), and by the EU H2020 framework program. His main research interests are Information Retrieval, Semantic Web, and Human Computation.