Computers and Education: Artificial Intelligence 5 (2023) 100151

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation

Hassan Khosravi [a,*], Paul Denny [b], Steven Moore [c], John Stamper [c]

[a] The University of Queensland, Brisbane, Australia
[b] University of Auckland, New Zealand
[c] Carnegie Mellon University, United States of America

ARTICLE INFO

Keywords:
Learnersourcing
Crowdsourcing in education
Student generating content
Human-AI partnership

ABSTRACT

Engaging students in creating novel content, also referred to as learnersourcing, is increasingly recognised as an effective approach to promoting higher-order learning, deeply engaging students with course material and developing large repositories of content suitable for personalised learning. Despite these benefits, some common concerns and criticisms are associated with learnersourcing (e.g., the quality of resources created by students, challenges in incentivising engagement and lack of availability of reliable learnersourcing systems), which have limited its adoption. This paper presents a framework that considers the existing learnersourcing literature, the latest insights from the learning sciences and advances in AI to offer promising future directions for developing learnersourcing systems. The framework is designed around important questions and human-AI partnerships relating to four key aspects: (1) creating novel content, (2) evaluating the quality of the created content, (3) utilising learnersourced contributions of students and (4) enabling instructors to support students in the learnersourcing process. We then present two comprehensive case studies that illustrate the application of the proposed framework in relation to two existing popular learnersourcing systems.

1. Introduction

Our increasingly connected world is empowering learners and enabling exciting new pedagogies. In particular, educational tools that facilitate collaboration between students can help to foster a wide range of social and domain-specific skills (Jeong et al., 2019). The literature on computer supported collaborative learning documents a diverse range of pedagogies that have been applied for decades in many subject domains and educational levels (Lehtinen et al., 1999, Roberts, 2005, Kaliisa et al., 2022). One recent approach, derived from foundational work on contributing student pedagogies (Collis & Moonen, 2002, Hamer et al., 2012), involves students creating and sharing learning resources with one another. Such activities have gained popularity in recent years and are associated with two broad types of benefits. Firstly, creating learning content is a cognitively demanding task that requires students to engage deeply with course concepts and exhibit behaviours at the highest level of Bloom's taxonomy of educational objectives (Hilton et al., 2022). Secondly, leveraging the creative power of many students can result in the rapid and cost-effective creation of large repositories of learning resources that can, in turn, be used for practice and to support personalised learning experiences (Singh et al., 2021).

Learnersourcing is a commonly used term to describe the practice of having students work collaboratively to generate shared learning resources (Kim, 2015). It is related to the more general task of crowdsourcing, in which tasks are outsourced to a pool of participants, often drawn from large and undefined populations, each of whom makes a small contribution to some product. The free online encyclopedia, Wikipedia,[1] is perhaps the canonical example of a crowdsourcing project where the number of users of the resource vastly outweighs the number contributors (Antin, 2011). Crowdsourcing participants are also rarely end users in the context of tools such as Amazon's Mechanical Turk, which is a platform that harnesses human computation in the form of microtasks to solve larger problems (Doroudi et al., 2018). In such cases, crowdworkers are typically paid a small fee for their contributions, although these kinds of models have drawn criticism around

* Corresponding author.
E-mail address: h.khosravi@uq.edu.au (H. Khosravi).
URL: https://hassan-khosravi.net (H. Khosravi).
[1] https://en.wikipedia.org.

https://doi.org/10.1016/j.caeai.2023.100151
Received 17 October 2022; Received in revised form 18 June 2023; Accepted 19 June 2023
Available online 24 June 2023

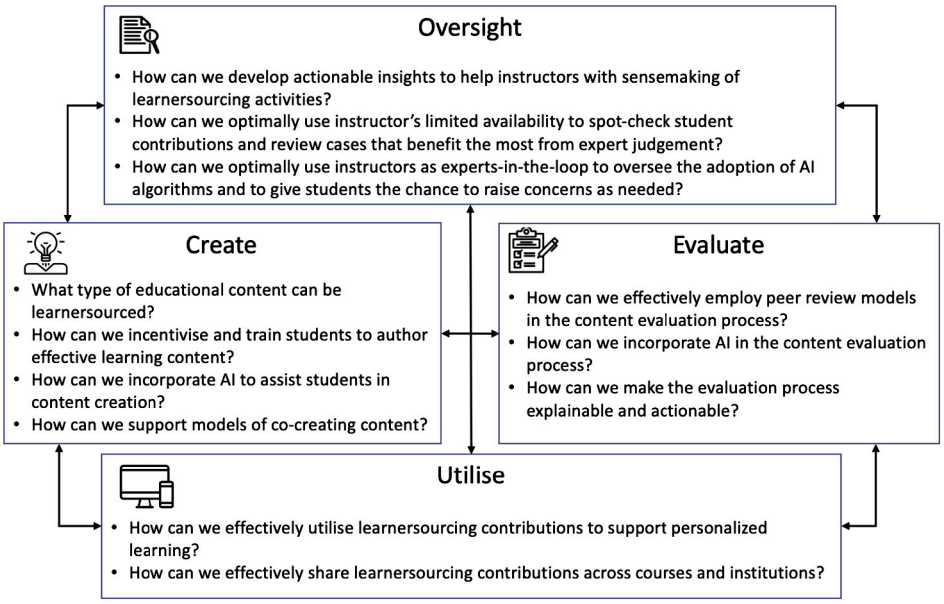


Fig. 1. A learnersourcing framework related to creating novel content, evaluating the quality of the created content, utilising learnersourced contributions of students and enabling instructors to support students in the learnersourcing process.

their exploitative nature (Schmidt, 2013). In contrast, learnersourcing adopts a more human-centred focus and involves a cohort of learners studying a common subject. The process of developing material in a learnersourcing activity is pedagogically beneficial to the learner, whereas in a more traditional crowdsourcing task it is a means to an end (Jiang et al., 2018). Learners are also inherently motivated to use the materials developed by their peers, and this is a defining feature of such pedagogies (Hamer et al., 2008).

As interest in learnersourcing grows, recent work has emerged seeking to inform the design of learnersourcing tools and to guide research activities. Khosravi, Demartini, et al. (2021) reflect on the experience of building an adaptive learnersourcing platform and present a series of data-driven lessons for developers and researchers. They recommend the use of accurate and explainable consensus approaches for assessing content quality, incentives for encouraging high-quality contributions and open learner modelling to make progress visible to learners. They also argue strongly for the need to harness the potential of artificial intelligence (AI) for improving feedback and generating effective recommendations for learners. More recently, Singh et al. (2022) propose a theoretical framework for studying and designing learnersourcing systems. Their framework is centred around a set of four design questions previously used to classify crowdsourcing systems, which they augment with two questions that focus on the prerequisite skills and the learning outcomes of those who contribute new learning artefacts. They apply their framework to classify prior learnersourcing literature, and offer it as a guide for practitioners when designing novel systems. They suggest that the two primary questions in their framework, "what is being done?" and "what are contributors learning from the task?", should be answered as the first step in any new initiatives. When considering future directions, they highlight the complementary relationship between learnersourcing and AI. For example, when meaningful learning activities naturally produce data that can be used for training models, which can then be used to evaluate resource quality and make personalised recommendations.

In the current paper, we propose a novel framework that captures what we see as the four essential components of learnersourcing models. Our framework, which is organised around fundamental *activities*, is complementary to that proposed by Singh et al. (2022) which uses a learner-centric rather than an activity-centric lens. They adapt the 'what', 'who', 'why', and 'how' design questions proposed by Geiger et al. (2011), providing an excellent framework for classifying prior work. In contrast, our framework considers the creation, evaluation, utilisation and instructor oversight of resources to be the four defining activities in the learnersourcing model. In our view, one of the benefits of this activity-centric model is that each component of the framework represents a core activity in which human effort is directly enhanced through partnership with AI. Given the rapid and transformative emergence of generative AI powered by large language models (Kasneci et al., 2023), human-AI partnerships will play an essential role in the near future and thus a suitable framework to capture and describe these partnerships is needed. Prior learner-centric frameworks that focus on issues of learner mindset and motivation are very valuable, for example to categorise literature and design incentive systems, but are less suitable for capturing the increasingly important role of generative AI (Chao et al., 2017). In presenting our framework, we highlight several questions of interest for each component, and we discuss related work, current challenges and promising future directions.

We present the framework in Section 2, organising key questions around the 'Create', 'Evaluate', 'Utilise' and 'Oversight' components. Section 3 then presents two case studies that illustrate the application of the framework in the context of two existing learnersourcing systems; namely, PeerWise (Denny et al., 2008) and RiPPLE (Khosravi et al., 2019). These case studies are complementary as PeerWise is considered a pioneering learnersourcing system, originally developed in 2008. In contrast, RiPPLE was originally released in 2019 and is still under active development, which has enabled it to benefit from state-of-the-art insights from the fields of crowdsourcing, AI and human-computer interaction (HCI). Finally, Section 5 provides concluding remarks about the future of learnersourcing. In particular it highlights the growing impact of AI including the use of generative language models for creating new content, deep learning models for evaluating content quality, recommendation algorithms for helping students better utilise content, and the use of experts in the loop across the four dimensions of our framework.

2. Developing learnersourcing systems

This section presents the framework that considers related work, current challenges and promising future directions around the four key activities supported by learnersourcing systems. The framework is depicted in Fig. 1, and we organise the remainder of this section around these four components. Section 2.1 focuses on the authorship and cre-

ation of novel content by learners. Section 2.2 focuses on how the quality of learnersourced contributions can be effectively evaluated. Section 2.3 focuses on how repositories of content created via learnersourcing can be effectively utilised to support student learning. Finally, Section 2.4 focuses on the role of educators and how they can oversee and provide support to students engaged in learnersourcing.

As depicted in the framework in Fig. 1, each component interacts with the other components in various ways. When content is created by learners, that content can be evaluated by other learners, utilised for study and review purposes, and overseen by instructors with expertise in the subject area. In all cases, the feedback that is generated – either explicitly through evaluation or instructor oversight, or implicitly through utilisation – can be used by the author of the content to edit or improve their existing artefacts, or to create more effective content in the future. Insights generated through both the utilisation and evaluation of resources by learners can feed directly into the process of instructor oversight, for example to facilitate efficient use of the instructor's time to make high impact decisions. Similarly, aspects of expert judgement can be used to provide feedback to students when evaluating content and to utilise resources that are approved by the instructor. Finally, learner evaluation of content can feed directly into recommendation algorithms that improve the utilisation of that content. Conversely, utilisation patterns – for example, aggregated data on selected options or popular answers – can be useful data for learners when evaluating learnersourced content.

One of the key contributions of this paper is that it leverages recent developments in the fields of human-centred AI and AI in education to highlight opportunities for human-AI partnership across the four dimensions of the framework. In particular, it leverages recent developments from the field of AI in education such as the use of generative language models for creating educational content, deep learning models for evaluating educational content quality, learner modelling and recommendation algorithms for helping students better utilise content, and relying on instructors as experts-in-the-loop in the context of learnersourcing. We discuss these in more depth below:

- The creation of educational content is a vital aspect of learnersourcing. With respect to 'Create', our framework focuses on the potential range of content that can be produced, the methods of incentivising its creation, and the collaboration of learners and AI to jointly create resources. For example, generative AI models have proven adept at producing certain kinds of learning resources (Leinonen et al., 2023), and these models can be directly integrated into learnersourcing systems to improve the efficiency and the quality of the content that learners ultimately produce. We provide a concrete example of such integration, that of generating distractors for multiple-choice questions, as part of one of our case studies in Section 3.1.1.
- Evaluating the quality of learner-generated content is an essential aspect of learnersourcing systems. Although prior work has shown that much of the content generated by students is sufficiently high quality to benefit learning (Kelley et al., 2019, Walsh et al., 2018), the presence of low quality content negatively affects learning efficiency (Moore, Stamper, et al., 2021). With respect to 'Evaluate', we consider the use of peer review, AI assistance for improving the reliability of a peer review process and the provision of explainable and actionable feedback to content authors.
- As students utilise content, fine-grained interaction data can be collected in real-time and provided as input to learner models that produce personalised recommendations of content, improving learning efficiency (Papanikolaou, 2014). With respect to 'Utilise', we consider the application of such learner models as adaptive engines that recommend resources, as well as the use of learnersourced content outside of institutional silos.
- Manual review of learner-generated content by experts does not scale, and so support for efficient moderation is critical (Moore, Nguyen, Stamper, 2023). With respect to 'Oversight', we consider the provision of AI-assisted actionable insights to instructors, and the efficient use of their time to balance the cost and reliability of expert judgements.

2.1. Creating learnersourced content

The creation of educationally-relevant materials by students is central to learnersourcing activities, and the term *student-generated content* (SGC) is often used to describe the resources produced in the context of learnersourcing (Snowball & McKenna, 2017, Wheeler et al., 2008, Hardy et al., 2014). Creating content can be a challenging task for students, but one which offers several distinct learning benefits (Singh et al., 2021). Consider, for example, some of the typical processes that a student might follow when creating content as part of a learnersourcing activity. To prepare for creating a specific learning artefact, a student would benefit from studying the requisite material in order to understand the concepts being targeted. While constructing the artefact, the student may generate worked solutions or explanations of the relevant ideas, either for explicit inclusion as part of the material to be published or for their own benefit. Either way, such explanations are beneficial and research into the self-explanation effect suggests that students who explain examples to themselves, whether prompted or not, learn more effectively (VanLehn et al., 1992, Bisra et al., 2018).

Students also benefit from the generative aspects of content creation. When compared to reading content produced by others, which tends to be a passive activity, generating content leads to more robust recall (Crutcher & Healy, 1989). This phenomenon is related to the widely studied 'generation effect' which suggests that people remember information better when they take an active role in its production (Slamecka & Graf, 1978). Originally established within the context of simple memorisation tasks, the generation effect has been shown to generalise to more complex learning materials (Rittle-Johnson & Kmicikewycz, 2008, Kinjo & Snodgrass, 2000, Kelley et al., 2019) and across a variety of domains (DeWinstanley & Bjork, 2004, Scapin, 1982). More broadly, the content creation aspect of learnersourcing is an inherently active task, and more effective for building knowledge than activities that are more passive such as listening to lectures delivered by an expert. This active construction of knowledge is a central tenet of constructivist learning theory (Bada & Olusegun, 2015), which provides theoretical support for activities such as learnersourcing. Moreover, social constructivism posits that co-constructing knowledge with others provides an important cultural context on which personal knowledge can be built (Rannikmäe et al., 2020). In a similar vein, Bredow et al. (2021) cite constructivism and social constructivism as providing theoretical support for the academic and interpersonal benefits of flipped learning.

Despite these well established benefits, there remains some debate as to how to best incentivise students to create learning materials (Khan et al., 2020). For many educational activities, a common strategy to promote participation is to reward students with some form of mark or course credit. There is a tension here, as associating a large amount of course credit with a learnersourcing task places a burden of evaluation on instructors, whereas too small a credit can lead to resentment from some students that the reward is not commensurate with the effort (Doyle et al., 2019). Rather than making it compulsory, Singh et al. (2021) explored giving students the choice to create multiple-choice questions (MCQs) for their peers in the context of a large massive open online course (MOOC). They found that learners created higher quality content and valued the activity more when they could choose to participate rather than being required to do so, however fewer than 10% of learners voluntarily created content. While this may be suitable in the context of a large MOOC, in small class settings low levels of participation may limit the usefulness of the generated resource for practice purposes. From a user interface perspective, the use of virtual rewards such as points and badges has shown some promise for incentivising

students in learnersourcing contexts, however they tend to be more effective motivators for the utilisation rather than the creation of content (Yeckehzaare et al., 2020). For example, one study found that a badge-based achievement system had a significant effect on the number of questions answered by students for practice, but not on the number of questions authored (Denny, 2013). Subsequent work went further to establish a causal link between these types of gamification mechanics and learning outcomes, mediated by practice testing behaviour (Denny et al., 2018).

2.1.1. Types of content

In general, the learnersourcing model is broad and places no boundaries on the type of content that students can produce. Any type of content that is associated with a course and typically produced by an instructor could be learnersourced by students. In practice, assessment and review materials such as questions and exercises form a popular category of learnersourced resources (Moore, Stamper, et al., 2022). Their popularity may be explained in two ways. Firstly, it leaves the production of core instructional material in the hands of experts. This not only helps to ensure that students build appropriate knowledge prior to generating materials that assess that knowledge, but it can be viewed as less controversial than having primary learning resources generated by non-experts (Hamer, 2006). Secondly, certain formats of assessment and review resources are very familiar to students, meaning they have plentiful examples to draw from. A good example of this is the widely popular multiple-choice question (MCQ) format, which is the most common type of artefact explored to date in the context of learnersourcing. Student-generated MCQs appear in tools such as RiPPLE (Khosravi et al., 2019), Quizzical (Riggs et al., 2020), UpGrade (Wang et al., 2019) and PeerWise (Denny et al., 2008). Other kinds of practice questions and exercises that have been explored include complex assessments on circuits and electronics (Mitros, 2015), Structured Query Language (SQL) practice exercises for database courses (Leinonen et al., 2020) and both small-scale (Denny et al., 2011) and large-scale programming problems (Pirttinen et al., 2018).

Although student-generated questions are common in the learnersourcing literature, a wide variety of other instructional content has also been explored. For example, Gehringer et al. (2006) describe the use of their Expertiza tool for managing student-generated contributions to their course textbook, which included students making improvements to existing explanations contained in the book and creating new examples for the concepts described in each chapter. Other examples that illustrate the variety of SGC in the literature on learnersourcing include subgoal labels for video tutorials (Kim et al., 2013, Weir et al., 2015), subgoal hierarchies for programming exercises (Jin & Kim, 2022), the underlying knowledge components for assessment items (Moore et al., 2020), personalised hints for engineering design problems (Glassman et al., 2016), explanations for peer instruction questions (Bhatnagar et al., 2020), solutions to open-ended questions (Wang et al., 2019) and explanations for programming misconceptions (Guo et al., 2020). Hills (2015) describe the use of both learnersourced blogs and videos in a psychology course, where students generate content that aligns with their personal interests. Students were tasked with collecting existing resources from their everyday experiences and curating them on a blog, as well as producing novel content in the form of persuasive videos promoting pro-social messages.

2.1.2. Promoting high quality content

In order to create high quality content within a learnersourcing task, students need both domain-specific knowledge and task-specific knowledge (Devine & Kozlowski, 1995). The former is typically developed through engagement with course learning materials and other curricula resources. Alongside this core disciplinary knowledge, students also need to understand how to construct high quality learning resources. Depending on the type of artefacts being learnersourced, this may include knowledge of how to construct effective MCQs, or how to generate useful hints or helpful explanations (Snow et al., 2019). This task-specific knowledge may be taught directly by the instructor or incorporated as part of the learnersourcing system (Doyle et al., 2019).

Various approaches for the instruction of task-specific knowledge have been reported. Doyle et al. (2019) describe a learnersourcing activity involving MCQs where guides for constructing effective MCQs were made available to students on the course learning platform. However, they observed that these guides were infrequently consulted by students, despite the fact that students complained of the need for more support on how to construct questions. As a result, they recommend teaching the principles of good MCQ design explicitly, and providing examples of both good and weak questions to illustrate these principles. Bates et al. (2014) describe a deliberate approach to prepare students for generating MCQs, involving a 90 minute tutorial consisting of several elements. These included a content-neutral quiz to familiarise students with the terminology of an MCQ (e.g., stem, options, distractors), a self-diagnosis quiz to guide students towards a learning-orientation rather than being performance focused, a representation of Vygotsky's "Zone of Proximal Development" (Chaiklin, 2003) to challenge students to author questions of high cognitive value, and question exemplars. To assess the value in this deliberate approach to teaching task-specific knowledge, the authors evaluated the questions authored by students finding that 75% passed a set of quality criteria including explanation detail, distractor plausibility, question clarity and cognitive level. An even more rigorous approach to scaffolding the MCQ creation process was documented by Hilton et al. (2022), in which students were introduced to increasingly complex tasks in five distinct steps over a period of 10 weeks. These began with students constructing short statements that are either true or false, to practise writing concise statements targeting a single topic, and then suggesting improvements to MCQs they found online before progressing to authoring their own MCQs. The authors see their scaffolding of task-specific skills as essential to the broad and deep conceptual benefits they observed as a result of the learnersourcing task.

One commonly reported challenge with the learnersourcing of assessment items such as MCQs is that many of them end up being simple recall questions (Moore, Nguyen, et al., 2021). For example, in a study by Bottomley and Denny that used PeerWise for learnersourcing MCQs, more than half of the student-generated questions were classified at the lowest level of the revised Bloom's taxonomy (Krathwohl, 2002). In their study, students were provided exemplar questions but were not explicitly instructed on the learning objectives or the cognitive processes associated with them as described by a taxonomy like Bloom's. The extent to which such instruction is helpful is not clear and is likely highly contextual. In some situations, the use of exemplars alone has proven effective for training novices in other crowdsourcing contexts. For example, Doroudi et al. (2016) explored the effects of different training strategies on novices in a crowdsourcing task, finding that the provision of expert examples outperformed other training strategies. Similarly, in the context of learnersourcing subgoal labels, Choi et al. (2022) present learners with good examples of subgoal labels in a 'warm-up' training phase before they are asked to create their own labels. Also in the learnersourcing literature, Huang et al. (2021) reported good success explicitly teaching students about Bloom's taxonomy and its application to assessment items. Students practised assessing questions according to Bloom's taxonomy before creating their own, and the authors found that a selection of the student-generated questions performed as well as questions generated by academics on summative exams.

Moving beyond task-specific knowledge, Lahza et al. (2022) explore the benefits in a learnersourcing context of scaffolding self-regulated learning behaviours. In a controlled study, they investigated the benefits of explicitly prompting students to plan their work before creation, self-monitor during creation, and self-assess after creation. Although these metacognitive scaffolds have robust theoretical benefits, in practice the authors found that they increased task complexity and completion time, without any significant improvement in the quality of the content

students produced. In general, the extent to which different types of training resources are effective, and the tradeoffs they present in terms of time and scalability of instruction, is currently under-explored in the learnersourcing literature.

2.1.3. AI assistance in content creation

Creating high-quality and novel content is demanding, and not all students engage well with the generative aspects of learnersourcing (Moore, Nguyen, et al., 2021). Indeed, prior research has shown that students are often more inclined to use and evaluate resources that are created by others rather than expend the effort needed to create high quality content of their own (Singh et al., 2021, Pirttinen & Leinonen, 2022). Using machine learning techniques to automatically generate novel content, at least in a draft form that a student could refine, is therefore a promising approach for scaffolding the learnersourcing process with AI.

Large language models (LLMs) have recently emerged and proven very effective at generating realistic human-like content. Models like GPT-4 (OpenAI, 2023) and Open AI's Codex (Chen et al., 2021), which are respectively fine-tuned to produce natural language text and source code, have received a great deal of attention. Such models are not limited to textual output, with models such as DALL·E (Ramesh et al., 2022) being able to produce extremely creative artistic images from natural language prompts. These models tend to be very good at "few shot" learning, in which the input prompt includes one or more contextual examples which leads to the generation of a novel output. Early evidence suggests that these models are very good at creating educational content (Wang et al., 2022). Recent work by Drori et al. (2021) applied both GPT-3 and Codex to generate novel university-level mathematics problems with explanations, and to solve them with equivalent success rates to humans.

In natural language processing and machine learning, it is common to use questions to train models to both generate higher quality questions and enable them to answer them with higher accuracy (Wang, Manning, et al., 2021). Naturally, learnersourcing could provide a great source of training data for these applications. Such models may be useful in assisting with the content creation phase of learnersourcing. For example, a learner could provide an initial prompt to the model, which may include examples and other contextual priming information, and would then be able to evaluate and refine the content produced by the model. Such an approach places a greater emphasis on content evaluation than on content creation, and may improve the efficiency with which large-scale resources can be produced. (Sarsa et al., 2022) use the term "robosourcing" to describe this augmentation to the traditional learnersourcing model. In general, we expect that as generative AI models improve and become more deeply embedded in educational contexts, there will be a shift in emphasis with respect to the application of higher-order thinking skills. Creating novel content, traditionally seen as an important high-level skill, will become relatively less important when compared to evaluating and critically analysing existing content, given the ease with which AI models can produce new content. Similarly, the importance of developing certain lower-order, foundational skills may become relatively less important given the availability of AI models that provide suitable support (Denny et al., 2022).

2.1.4. Co-creation models

Most existing learnersourcing systems expect individual learners to generate and contribute complete learning resources (for example, in the case of MCQs, this would typically include the stem, a set of distractors, and an explanation or sample solution) (Singh et al., 2022). This requires a large investment of time from a learner and can be error prone as individual parts of the artefact are not reviewed by others before being assembled. In contrast, many crowdsourcing systems tend to be structured around small micro-tasks (e.g., labelling one image (Chang et al., 2017)) which permit useful contributions with little time and effort. Moreover, these systems often support a co-creation model where individuals can collaborate on the same artefact. Examples of this type of model include editors in Wikipedia being able to directly modify content produced by others, and forums like StackOverflow allowing users with sufficient reputation to edit questions and responses generated by other users. Another early example of this idea was the novel wordprocessor plugin by Bernstein et al. (2010), called Soylent, which permitted crowdworkers to make small proof-reading edits to a document, including minor formatting and wording changes.

Learnersourcing can make use of co-creation models to empower individual learners to contribute in a variety of different ways. Singh et al. (2021) explored the factors that influenced learners' decisions to create content when it was optional to do so. Lack of time, low confidence and lack of interest were the primary reasons cited for choosing to not create content. As a result, they advocate for co-creation models that allow for tiered contributions such that students with little time or less confidence could contribute in more granular ways. In the context of their study, which involved authoring MCQs, they suggested that rather creating a complete MCQ, learners could provide a set of distractors or write an explanation. Recent work by Kim et al. (2022) explored this very idea through a modularised approach to learnersourcing MCQs. In their model, individual components of a question, such as a single option or a stem, could be authored and refined by learners thus providing flexibility for different learners to contribute in ways that suit their interests.

2.2. Evaluating learnersourced content

One of the effects of learnersourcing is that it makes it relatively easy to develop large repositories of SGC. While strong evidence from previous work suggests that a large portion of the SGC is of a high-quality and meets rigorous judgemental and statistical criteria (Walsh et al., 2018, Galloway & Burns, 2015), it also suggests that students commonly create resources that are ineffective, inappropriate, or incorrect (Tackett et al., 2018, Denny et al., 2009, Bates et al., 2014). As a consequence, to effectively use SGC repositories, there is a need for separating high-quality from low-quality resources. One approach is to engage instructors as experts in evaluating the quality of the resources; however, the instructor-led quality evaluation is not scalable and can be expensive due to the potentially large size of the repositories (Section 2.4 explores plausible methods of optimally using instructor's limited availability towards evaluating content). This section explores two alternative approaches of employing human or machine computation for evaluating the quality of the resources. Primarily, we first explore the possibility of incorporating co-regulation models of a peer review process where students are engaged in assessing the quality of resources authored by their peers. We then explore how AI methods can be incorporated to help with the assessment of the quality of the resources. We finally discuss the need for accurate, fair and transparent evaluation methods, regardless of whether they are done by humans or machines.

2.2.1. Peer review models for evaluating quality

Peer review is a well established model for evaluating quality, often employed for academic publishing (Tennant, 2018). To determine the suitability of the peer review process in the context of learnersourcing, it is important to first consider whether engaging students in evaluating content are beneficial to their learning. If such benefits are absent, ethical issues arise with respect to utilising students as cheap labour to reduce the workload of instructors (Zdravkova, 2020). There is however a general consensus that engaging students in peer review has many benefits (Nicol et al., 2014). These include enabling them to improve their comprehension of the content (Li et al., 2010), develop evaluative judgement (Tai et al., 2018, Gyamfi et al., 2021b, Khosravi, Gyamfi, et al., 2021) and a sense of accountability (Kao, 2013), and improve their writing (Polisda, 2017) and ability to provide constructive feedback (Lundstrom & Baker, 2009).

An equally important question to consider is whether students have the capacity to effectively evaluate the quality of peer-created resources. Prior work suggests that students by and large have the ability to accurately evaluate the quality of learning content (Galloway & Burns, 2015, Tackett et al., 2018). Having students as evaluators also addresses expert blind spot challenges as they would evaluate the effectiveness of a resource based on their own previous misconceptions (Nathan et al., 2001). However, as experts-in-training, their judgements cannot wholly be trusted (Abdi, Khosravi, Sadiq, Demartini, 2021). A common solution, which is also incorporated in academic publishing, is to rely on the wisdom of the crowd rather than one person by employing a redundancy-based strategy and assigning the same reviewing task to multiple users (Reily et al., 2009). Other strategies such as utilisation of rubrics (Gyamfi et al., 2021b, 2021a), exemplars (Carless et al., 2018), guides on providing effective feedback (Darvishi, Khosravi, Abdi, et al., 2022) and comparative judgement where students choose the 'better' of two pieces of work (Cambre et al., 2018, Palisse et al., 2021) have also been shown to be an effective method for helping students develop evaluative judgement.

The peer evaluation process can determine the way in which resources are shared. For example, one approach is to allow optional and subjective ratings on artefacts that have already been made available to all students (Denny et al., 2008). In such cases, aggregate ratings can be used to help students search for artefacts that have received higher quality scores (Denny et al., 2009) or be used to support personalisation mechanics (Williams et al., 2016). A more restrictive approach might follow the academic publishing process in which artefacts are judged by a subset of users using multi-criteria rubrics. In this case, reviews are used to determine whether a resource is of high enough quality to be approved and shared with other students or if it lacks the required quality and is to be rejected and sent back with feedback to the author (Khosravi et al., 2019). A workflow was developed by Lee et al. (2020) and deployed into a system known as Questionable, that allowed students to author and review questions that were then presented to their peers. Using this workflow, students would leave feedback regarding a question generated by their peers, indicating how it might be improved or assessing the quality and usefulness of it. These reviews were then presented to course staff, that used them to quickly determine the quality of the question or make any necessary changes to them. Ultimately this allowed the course to make quick use of the questions and ensure only the highest quality ones were being shown to the students.

2.2.2. AI in content evaluation

While the notion of following an academic publishing model for evaluating SGC is time-effective and supports student learning, it does introduce new challenges where AI can be of assistance. Here we provide two examples; firstly, unlike the publishing model where a meta reviewer makes the final call, in the case of SGC, it is impractical to expect the instructor to meta-review possibly thousands of artefacts that are being created in their course. Therefore, the process of deciding whether an artefact is to be approved or rejected based on multiple reviews needs to be automated. This raises a new problem commonly referred to as the consensus problem (Roitero et al., 2023). In the absence of ground truth, how can we optimally integrate the decisions made by multiple people towards an accurate final decision? Traditional consensus approaches rely on general statistical aggregations, such as the arithmetic mean, median, or majority vote (Zheng et al., 2017). However, previous studies have shown that there is an evidential difference in the judgemental ability of students (Abdi, Khosravi, Sadiq, Demartini, 2021). Inspired by work from crowdsourcing (Zheng et al., 2017), an interesting approach has been to use machine learning models to infer the reliability of a reviewer or a review such that the consensus model can put more weight on decisions made by reliable reviewers (Darvishi et al., 2021). Another challenge with relying on student reviews and feedback relates to students' failure in providing high-quality feedback, which leads to substantial negative consequences such as lowering standards (Yeager et al., 2014), reducing trust in the outcome (Carless, 2009), and making reviewees less likely to revise their work (Sommers, 1982). Here, inspired by advances in natural language processing to evaluate the quality of a review (Negi et al., 2016, Devlin et al., 2018), a possibility is to develop quality control functions that automatically assess the quality of the submitted feedback and ask students to improve, if necessary (Darvishi, Khosravi, Abdi, et al., 2022).

There have also been various attempts to automatically evaluate the quality of MCQs and more broadly educational artefacts. Metrics such as the discrimination index from classical test theory have been used for decades for identifying the quality of MCQs (Malau-Aduli et al., 2014). However, a limitation of this method is that it requires large quantities of data on student responses to items. Therefore, it cannot be used to evaluate the quality of new questions. The 2020 education challenge (Wang, Lamb, et al., 2021) from the Conference on Neural Information Processing Systems (NeurIPS) has started a new wave of using advanced AI models for the automatic determination of the quality of MCQs (Task 3 of the challenge). Three teams were announced as co-winners of this task, each achieving an 80% agreement with human evaluators' judgements. The approaches by two of the three winning teams Shinahara & Takehara and TAL Education presented solutions that computed explicitly-defined features based on the hypothesis that high-quality MCQs are appropriately difficult, readable, and have a balance among answer choices (Wang, Lamb, et al., 2021). Interestingly, the other winning approach by McBroom & Paassen did not use any complex feature engineering and had the very simple hypothesis that the quality of an MCQ correlates with the confidence of students answering it (McBroom & Paassen, 2020). They argued that high student confidence implies that the question is clear and unambiguous. In addition, they argued that the Dunning-Kruger effect (Dunning, 2011) may result in students holding key misconceptions by reporting high confidence in incorrect answers if the question clearly addresses this misconception. More recently, Ni et al. (2021) propose DeepQR that, alongside computing explicitly-defined features, uses a 2-layer transformer encoder to consider semantic features, which are designed to capture relations between different question components. Compared to six existing models, DeepQR was able to more accurately identify questions that were low or high quality. Another study trained a state-of-the-art language model, GPT-3, on learnersourced questions to classify the quality of a question as low or high and the cognitive level, according to Bloom's revised taxonomy (Krathwohl, 2002, Moore, Nguyen, Bier, et al., 2022). They then had the model classify student-generated short answer questions to automatically classify their quality and cognitive level.

It is worth highlighting that the majority of existing work has focused on automatic evaluation of MCQs (Kurdi et al., 2020). However, the advancements in natural language processing (NLP) on pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and its extended models make it possible to generate and automatically detect the quality of content, which can also be tailored toward educational artefacts. Another limitation of the current AI-based evaluation methods is that they focus mostly on explicitly-defined and semantic features rather than the correctness of the content. An interesting future direction is studying how AI and students can collaborate on content evaluation where the correctness of the content is examined by students and the readability and flow is examined by AI.

2.2.3. Explainable and actionable evaluation methods

Much of the existing work on evaluating learnersourced content has focused on just separating out high and low quality resources (Abdi, Khosravi, Sadiq, Demartini, 2021). However, an important aspect of engaging students in learnersourcing and evaluating their work is to help students not only improve their ability in content creation but also to help them develop their ability to monitor, evaluate and regulate their learning so that they can revise and enhance their created content. This

is of particular importance in the cases where learnersourcing activities are tied to assessment in which the assessed quality of a resource impacts student grades (Singh et al., 2021). In the case of using peer review, an interesting future direction is to study how best practices from feedback literacy in terms of engaging authors and evaluators in dialogue (Ajjawi & Boud, 2017) and feedback loops (Carless, 2019) can be applied in learnersourcing. In the case of AI-based evaluation, the current use of deep learning methods for assessing the quality of educational artefacts such as MCQs has shown to be accurate and close to human judgement; however, the models operate as black boxes, providing no justification for their decisions. An interesting future direction is to study how explainable AI methods in the context of education (Khosravi et al., 2022) can be applied to learnersourcing.

2.3. Utilising learnersourced content

A side benefit of engaging students in content creation and evaluation is that it enables the development of large repositories of learning resources, which can be shared with students to provide practice opportunities (Singh et al., 2022). For instance, a plethora of studies describe the generation of large question banks across a variety of domains, and the production of millions of MCQs, that have been subsequently used for practice purposes Moore, Stamper, et al. (2022). However, the utility of the developed repositories has often been limited to the courses where the content was originally produced, which restricts the life and impact of the created content. In addition, typically students only have access to simple search and filtering functionality for the selection of practice questions. This may lead to students spending their time ineffectively on resources that are targeted toward an average student of the class rather than focusing on their knowledge gaps (Koedinger et al., 2013). This section explores how these limitations can be addressed so that SGC repositories can be more effectively utilised. We first discuss approaches for supporting personalisation in engaging with SGC repositories. Then we discuss how SGC repositories can have an impact beyond the course of their origin and to be utilised in future offerings of different courses across institutions.

2.3.1. Personalisation in engaging with content repositories

Learnersourcing yields student-generated artefacts, such as questions, but the process of students interacting with such learnersourcing activities and systems also generates auxiliary user-item-outcome data that can be leveraged for developing learner models (Abdi, 2022). A learner model is an abstract representation of a student's knowledge state. They are a core component of adaptive educational systems that provide students with customised learning paths and adaptive feedback based on their learning process (Koedinger et al., 2013). In the case of modelling learners in learnersourcing, Moore, Nguyen, Stamper (2022b) and Abdi et al. (2019) have utilised data collected from students' interactions with a repository of SGC to model the knowledge and skills required to solve problems in the context of chemistry, programming and relational databases courses. As the process of creating and evaluating content also leads to student learning, data captured during the learnersourcing activities can also be incorporated into these learner models akin to how student data from problem attempts is currently utilised. Empirical results from two studies that present learner models, which utilise data from students' learnersourcing tasks, demonstrate that these models outperform learner models that only utilise traditional assessment data (Abdi, Khosravi, Sadiq, 2020, Khosravi, Demartini, et al., 2021). Educational recommender systems can make use of a learner model to recommend learning content, which can optimise student learning and their time spent. For example, Khosravi et al. (2017) present a new recommender system that recommends resources from an SGC repository. They found their approach to be able to adequately provide personalised recommendations for students who have previously used the platform as well as cold start users who are new. Abdi, Khosravi, Sadiq, Gasevic (2020) report that complementing recommendations on content from an SGC repository with a learner model lead to an increase in student engagement and a positive effect on students' perceptions of the quality of the recommendations.

Another way to support personalisation, which is commonly referred to as step-loop or inner-loop adaptivity, is to enable an adaptive instructional system to provide support to learners within a particular learning task (e.g., hints or explanations they receive) based on their performance (Aleven et al., 2016). A project by Glassman et al. (2016) designed a learnersourcing system to provide hints to students working through a college-level programming course. As students worked on the problem, they were able to automatically receive student-generated hints that would continually update or they could elect to only receive the hints when they requested them in a just-in-time fashion. This personalisation allowed students to leverage hints in a manner that fit their preferences. They proposed two models for their hints. The first is a push model, where the student-generated hints are presented to learners and constantly updated. The second is the pull model, where learners only receive hints when they request them. These two model approaches were extended by Singh et al. (2022) and applied to broader learnersourcing applications. They indicate that the push model can be leveraged when the learnersourced artefact is intended to help the learners in the problem-solving process. On the other hand, the pull model should be utilised when the learners have completed a problem, but they might be seeking a more optimal solution. Another related system known as SolveDeep leverages student-generated sub-goals on algebra problems to provide feedback to other learners' solutions, akin to providing a hint (Jin et al., 2019). They found that participants effectively leveraged the subgoals generated by other learners, which helped them effectively and efficiently solve several algebra problems. Another example of supporting inner-loop adaptivity in learnersourcing comes from the popular Axis system (Williams et al., 2016), which enables students to generate explanations for math problems and then used adaptive multi-armed bandit algorithms to deploy the optimal explanations to students. They found that explanations delivered to students led to higher learning gains than a majority of the existing explanations previously used for the problems.

2.3.2. Sharing learnersourcing contributions

A present challenge in the learnersourcing space is the sharing of SGC across courses, institutions, and ultimately to a broader audience. While efforts in the open educational resources (OER) space provide insights into the dissemination of instructional and assessment content, it is often content that is created by professional instructors that intend to share their content (Wiley et al., 2014). Potential issues around copyright and the leaking of question bank answers create challenges on how we can readily share these materials in a way where they can effectively be used for both formative and summative assessments of student learning. Platforms such as OpenStax and ASSISSTments, popular OER platforms, attempt to address such challenges by requiring instructors to verify their identity before accessing the materials (Pitt, 2015, Heffernan & Heffernan, 2014). Researchers and practitioners continue to expand their efforts in sharing learnersourcing contributions from their courses and systems. For example, Quintana et al. (2018) have students in a data science course develop questions that were then utilised by students in future semesters of the same course. In the context of MOOCs, (Kim et al., 2014) and (Weir et al., 2015) learnersourced labels for instructional videos that were leveraged by learners across multiple courses. (Kay et al., 2020) explore how students sharing resources and learnersourcing across multiple institutions can effectively be handled. Nevertheless, approaches and adoption of methods that support sharing learnersourced contributions are under-researched and -explored.

Even when the learnersourced content is created at a single institution, factors such as the student demographics and their location may impact how SGC is utilised (Moore, Nguyen, Bier, et al., 2023) Morales-Martinez et al. (2020) investigated how students at the same institution,

but split geographically between campuses in the United Kingdom and China, perceived the SGC of their peers. They found that when the students were identifiable, it had a significant impact on how their content was accessed and rated by their peers, such as students intentionally avoiding content created by classmates of certain nationalities. Other work in the space had more positive results, as Denny et al. (2012) had students in an introductory programming course at an institution in New Zealand generate learning resources for students in a similar course in Canada. The results indicated that this cross-institutional learnersourcing worked as well as it previously did within-institution and students from both institutions indicating that they prefer their contributions be shared more widely.

2.4. Overseeing the creation of learnersourced content

The role of the instructor in the learnersourcing process can vary depending on what system they might be using, how they elect to utilise the SGC in their course, or what they want to gain by having students participate in learnersourcing (Khosravi, Demartini, et al., 2021). No matter their role, they have a form of oversight in the learnersourcing process that enables them to gain insights and ultimately facilitate student learning. While in theory learnersourcing systems can operate without the presence of an instructor, academic oversight of the creation and evaluation process can serves as a demonstration of reliability, providing assurance to both educators and students that the system is trustworthy and dependable and to encourage high-quality contributions and peer reviews (Darvishi et al., 2023). It also provides useful insights into the student learning process that instructors can act on, such as modifying their curriculum based on where students have difficulty or having students create questions over a particular content area. Additionally, as the size of both in-person and online courses increases, resulting in increased student-generated contributions, it can be challenging for instructors to effectively use their time in evaluating and utilising these resources (Ji et al., 2022). While the role of the instructor is under-explored in the learnersourcing literature, work in the related fields of learning analytics, crowdsourcing, and machine learning can provide valuable insights that can be adopted in this context.

2.4.1. Learning analytics and actionable insights

Imagine a tool that is supporting engagement with learnersourcing for a class with over 500 students. How can the teaching team make sense of students' engagement and performance and how can they effectively facilitate learning? Data, in addition to the content, generated from learnersourcing activities can readily be leveraged by analytic systems and learning analytics dashboards (Matcha et al., 2019) with actionable insights (Jørnø & Gynther, 2018) to help instructors make sense of student learning and intervene with pedagogical interventions as necessary.

This raises the question of what metrics and analytics can be obtained from learnersourcing that instructors might find insightful. General metrics such as the number of logins, resources created, evaluated and attempted are readily available (see the case studies from Section 3) which might give instructors a sense of student engagement. Research has shown that student participation and completion are key indicators to increased learning (Moore, Nguyen, Stamper, 2022a). Studies also show that interaction with learnersourcing activities is not constant across student populations. The 90-9-1 rule was noted to apply in learnersourcing (Khosravi, Demartini, et al., 2021), stating that 90% of users are lurkers, 9% create some content, but the majority is created by 1% of the student population, which is similar to participation in forums within online courses. Engagement analytics may lead to instructor actions of congratulating high-achieving students and sending nudges and reminders to inactive students (Plak et al., 2023). Systems, such as RiPPLE, have also provided a type of learning analytics dashboard that can be leveraged by instructors to view questions students are struggling with or content areas where students may require additional support (Khosravi, Demartini, et al., 2021). Through the use of such analytics, students could then be encouraged to generate content in these troublesome areas, which could result in them thinking critically about the area, while also having the benefit of generating additional practice questions. In addition, performance data on students' interactions with assessment items can be leveraged to model student learning (Abdi, Khosravi, Sadiq, 2020, Kay et al., 2020) and to infer the quality of assessment items (Huang et al., 2021). More fine-grained data about the sequence of activities conducted by students can be leveraged to identify underlying tactics and strategies (Matcha et al., 2020) that are used by students while engaged in learnersourcing (Lahza et al., 2023). Another use case of actionable analytics for learnersouricing systems is to help instructors find resources or peer-review cases for being spot-checked which is discussed next.

2.4.2. Evaluating/spot-checking

Peer evaluation is an essential component of learnersourcing, as students commonly review the contributions of their peers within the same course or even across multiple institutions (Darvishi et al., 2021). However, peer evaluation is often susceptible to students being unmotivated to evaluate the work of their peers in a diligent manner (Liu & Chen, 2016). To make the peer review process more reliable, one potential approach is to utilise spot-checking, where an instructor or tutor evaluates some assignments and offers a reward to students who grade in a similarly diligent manner (Cambre et al., 2018). Various metrics may be utilised for optimally determining the resources which would benefit from spot checking the most. Gao et al. (2019) show that even random spot checking can incentivise reviewers to be more diligent. Wang et al. (2020) take a game theoretic approach to suggest optimal spot-checks to maximise the evaluation accuracy of reviews. Darvishi, Khosravi, Sadiq, et al. (2022) use a range of human-driven metrics (e.g., high-disagreement in moderation evaluations, a high ratio of downvotes in comparison to upvotes) and data-driven metrics (e.g., assessment items that have a low discrimination index or questionable distractors where the popular answer is not the one proposed by the author) in the context of learnersourcing to categorise resources into having high, medium, low or no priority for being reviewed.

It is worth noting that an alternative or complementary strategy to spot checking is to incorporate calibration submissions (Wang et al., 2020) for which the true grade is known. These submissions can be used for training where reviewers would have access to how an expert would have graded the task. A side benefit is that they can be used to infer the reliability of a reviewer. Relatedly, previous work by Hamer et al. (2005) has also explored a technical approach to this spot-checking problem, using algorithms to calibrate peer review scores automatically. Their approach identifies "rogue" reviewers that appear to assign their scores arbitrarily. Scores provided by these reviewers are weighted lower when the computed aggregate score for the artefact under review is determined.

2.4.3. Human-in-the-loop

When it comes to AI and machine learning, human judgement is still needed to make sense of or improve the results of the AI (Divate & Salgaonkar, 2017). This is one area where learnersourcing can be leveraged and the proposed framework can provide the insights to make it work. Human judgement and human-in-the-loop becomes even more important when we move away from traditional measures of educational achievement and focus on issues around fairness, accountability, transparency and ethics (or FATE), which was the focus of a recent special issue journal that had a common theme among the papers – that humans are still necessary "in the AI loop" to manage these issues (Woolf, 2022).

In a number of systems, AI guides the use of learnersourced content with human input. One example is AXIS (Williams et al., 2016), which uses multi-arm bandit methods to choose which student gener-

ated explanations to give students who need help. In this system, there is the opportunity for students to provide ratings on the explanations. These ratings and the future success of students on similar problems are combined to give the highest rated and best performing explanations a higher probability to be seen by future students.

One emerging area where human oversight may be needed in future learnersourcing systems is with the use of generative language models. Sarsa et al. (2022) propose the idea of robosourcing, where content generated by language models can be used as a starting point for students to accelerate the learnersourcing process. On the one hand, the increasing automation supported by such models may suggest less need for human input, but there is a need for caution. In their review of the opportunities and risks offered by foundation models, Bommasani et al. (2021) explicitly warn against the removal of teachers from the loop. Large language models are trained on broad data produced by humans, and thus are known to suffer from biases similar to humans (Chan, 2023). Using automatically generated content as the basis for learnersourcing tasks runs the risk of perpetuating some of these biases. We see a human-in-the-loop approach, involving both students and instructors, as essential for moderating such biases and for improving and tailoring the performance of the underlying generative models for suitability in learnersourcing contexts.

3. Case study

In this section, we present two comprehensive case studies that illustrate the application of our proposed framework. Section 3.1 discusses a pioneering learnersourcing system, PeerWise, and Section 3.2 describes a state-of-the-art learnersourcing system, RiPPLE, in the context of the proposed framework.

3.1. PeerWise

PeerWise (Denny et al., 2008) is a web-based learnersourcing tool, first developed in 2008, which supports students in creating, publishing, answering and discussing MCQs. As of the time of writing, approximately 7,000,000 questions have been published by students at 3,000 institutions worldwide.[2] Organisationally, the content within PeerWise is arranged hierarchically into "institutions" and "courses". Typically, an instructor would create a new course repository associated with their institution, and then grant their students access to that repository. Instructors and students use the same interface, although additional features are available for instructors to provide oversight, such as running basic usage reports and managing access permissions. Fig. 3 shows an example from the perspective of a student of the main menu for one course repository. Questions are organised with respect to whether they have been authored or answered by the student. The user interface of PeerWise keeps student identities anonymous, which is a deliberate design choice that has been shown to reduce certain kinds of biases in online learning environments (Morales-Martinez et al., 2020).

Instructors can also access fine-grained data for their courses, which includes timestamped records of all student interactions. The availability of this data has facilitated the work of educational researchers exploring various aspects of learnersourcing. To date, 123 articles by 262 distinct authors have been published using data collected by PeerWise.[3] Much of this work has focused on learning effects. For example, Kay, Hardy and Galloway used multilevel modelling to analyse data from 3,000 students over three years and across 18 physics, chemistry and biology courses at three UK universities (Kay et al., 2020). When controlling for prior ability, they found a significant positive association between students' engagement with PeerWise and their performance on end of course exams. They conclude that PeerWise offers a "low-risk"

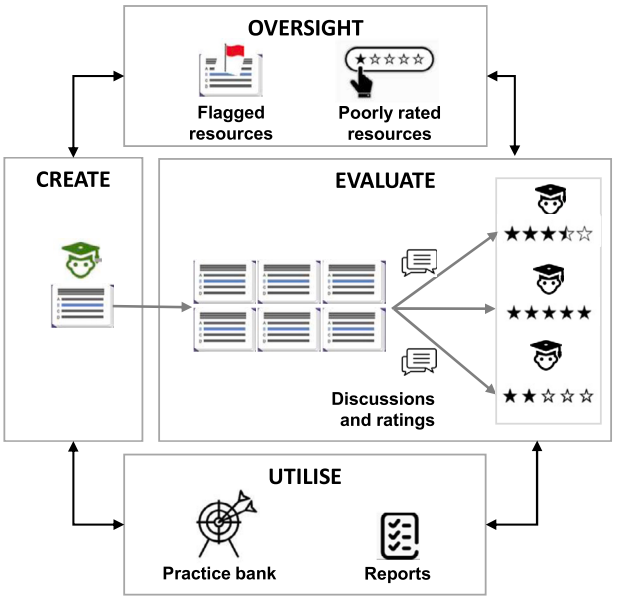


Fig. 2. Overview of PeerWise.

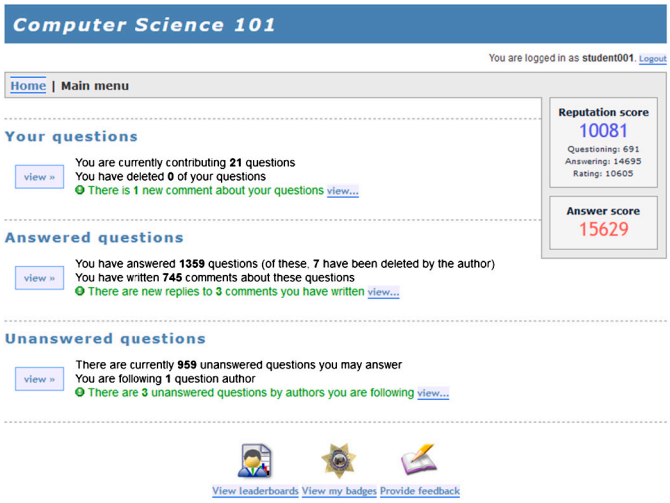


Fig. 3. Main menu of PeerWise.

and "low-cost" intervention that supports student learning and, more generally, that learnersourcing of course material in a structured way can provide measurable educational benefits.

Fig. 2 provides an overview of the operation of PeerWise with respect to our framework. We elaborate on each part of the framework in the following sections.

3.1.1. Create in PeerWise

The multiple-choice question (MCQ) format is simple, widely used in practice, and is a familiar format to most students. When constructing an MCQ in PeerWise, a student provides a *question stem* (a short section of text that describes the problem to be solved), a set of possible *answer options* (between two and five alternatives are allowed, exactly one of which must be selected as the correct answer) and an *explanation* (detailing the answer to the question and optionally explaining why certain alternative answers are incorrect). Fig. 5 illustrates an example of the stem, options and explanation for a question published by a student in an introductory MATLAB programming course.

As soon as a student publishes a question to the course repository, it is available to all other students in the course to answer and evaluate. That is, there is no formal validation or evaluation step prior to the

[2] peerwise.org.
[3] https://peerwise.cs.auckland.ac.nz/docs/publications/.

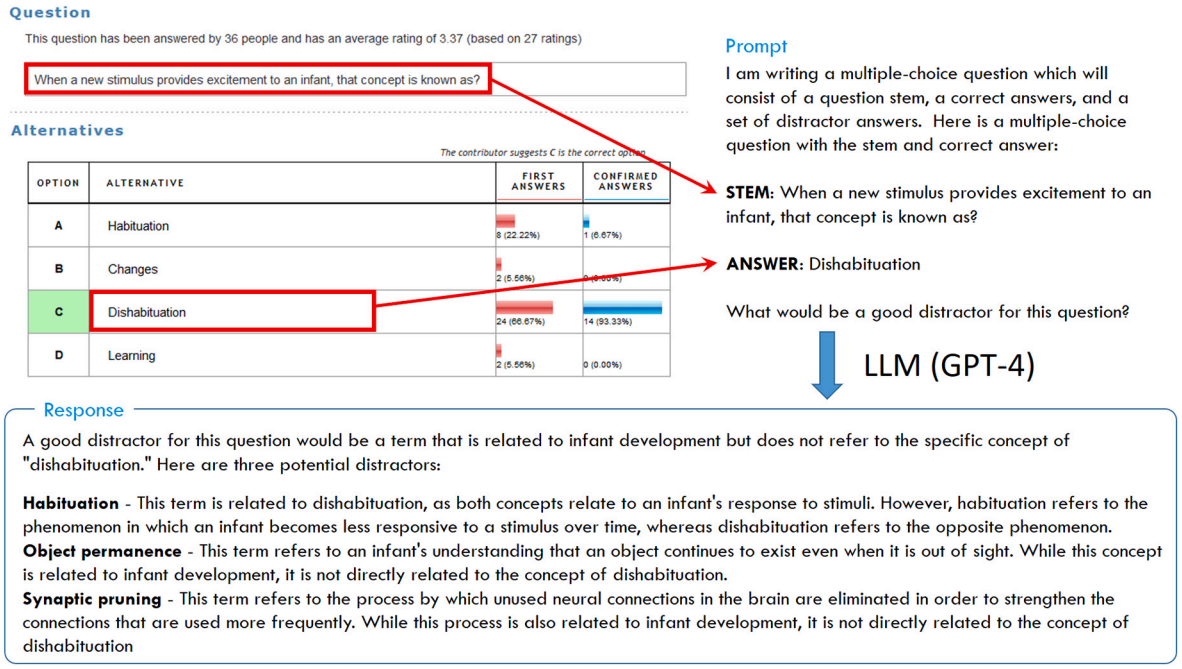


Fig. 4. Example of using a large language model to aid the creation process.

question being visible to other students. To assist students in finding questions of interest during utilisation of large question banks for practice, question authors are able to assign topic "tags" to their questions which can be used for searching the question bank.

One obvious human-AI partnership that can enhance the "create" activity in a learnersourcing environment is the use of a large language model, such as GPT-4, to provide automatic guidance to an author generating content. Consider, for example, a student using PeerWise to create a multiple-choice question as shown in the upper left portion of Fig. 4. As soon as the author of this question has crafted the question stem and indicated the correct answer, these two elements can be incorporated into a simple prompt (shown on the upper right of Fig. 4) and sent automatically to a large language model (in this example, GPT-4 is used). The response from the model, shown in the bottom portion of Fig. 4, presents three possible distractors for this question. These can then be presented to the student to help them craft an effective question. In this particular example, the question was already published and the responses selected by students utilising the question are illustrated by the red and blue histograms in the "First answers" and "Confirmed answers" columns, but naturally this data would not be available to the author at the time of writing the question. Note that the first distractor suggested by the language model, "habituation", actually turns out to be the most effective distractor in practice (i.e. the incorrect option selected by more than 20% of students attempting this question). The fact that this option was not provided to the language model as part of the prompt indicates that it is able to formulate suggestions that have utility. The other suggested distractors may have proven more effective, had they been used, than the current distractors which were rarely selected.

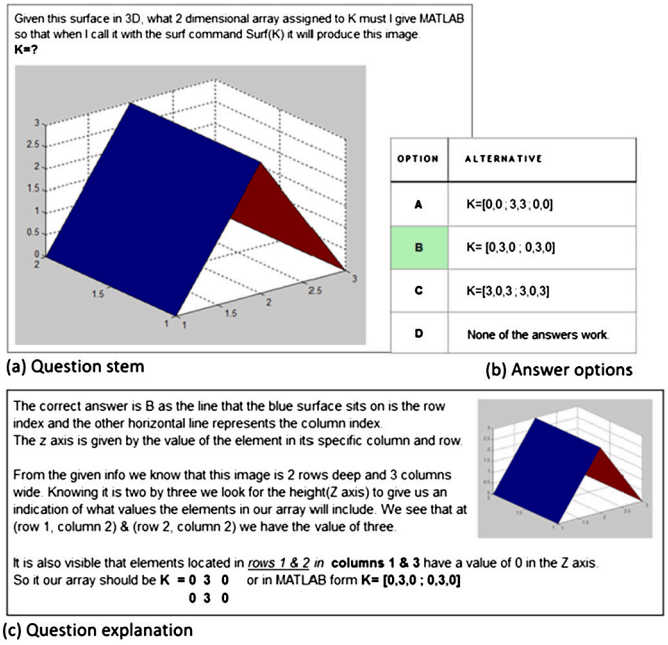


Fig. 5. Example of question stem, options and explanation.

3.1.2. Evaluate in PeerWise

Any student may evaluate any of the questions in the question bank, however they must first attempt the question by submitting an answer. Submitted answers are assessed through comparison with the question author's suggested answer, and the answers submitted to the question by other students. PeerWise generates one of seven possible feedback responses each time an answer is submitted. For example, if the submitted answer matches the author's suggested answer and that is also the most popular answer selected by other students, then the feedback response includes a solid green tick and a short descriptor message. An answer is deemed incorrect, and denoted by a solid red cross, if that answer differs from the author's suggested answer yet the suggested answer is the most popular answer selected by other students. Variations to this feedback, with appropriate descriptors, are shown when there is mixed agreement between the submitted, author's and most popular answer.

After submitting an answer and receiving this feedback, students are also shown the question author's explanation and a histogram of the options selected by other students. At this stage, a student can add a comment to the comment thread for the question, and they can evaluate the question by submitting a rating for its difficulty (3-point scale) and quality (6-point scale). There is also an option to "flag" any question

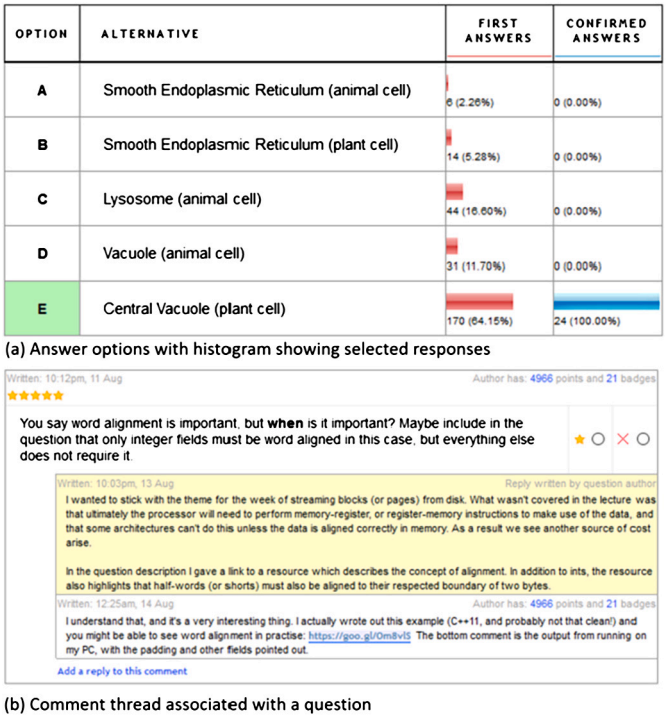


Fig. 6. Answer histogram showing distribution of selected responses and comment thread.

deemed inappropriate, which will bring it to the attention of the course instructor facilitating their oversight role. Finally, in light of seeing the feedback and discussion on the question, the student can optionally submit a "confirmed" answer which indicates the option they believe is correct. Alongside the ratings, response histograms and comments, these confirmed answers serve to improve the quality of the feedback students receive when utilising the resource for practice. Fig. 6 shows an example of the answer histogram for a question in a cellular biology course (with initial question attempts shown in red, and "confirmed" answers in blue), and a short excerpt from the comment thread to a question in a computer architecture course.

3.1.3. Utilise in PeerWise

The prior evaluation step provides two mechanisms for helping students locate high quality questions when utilising the repository for practice. Firstly, the difficulty and quality ratings are aggregated and can be used directly by a student to avoid low quality questions and to select questions at a suitable level of difficulty. Secondly, after evaluating a question a student can choose to "follow" the question author if they find the question was particularly helpful. Consistent with the anonymous interface design, students do not know the identities of the authors when choosing who to follow, and instead must make quality judgements on the basis of the content. Once following a particular author, a student gains access to all of the other questions that author has created in a separate section of the tool. As a way of incentivising the authoring of useful questions, students can see how many of their peers are following them.

A common use of question repositories in PeerWise is for review and practice purposes leading up to summative tests and exams. Prior work has shown that answering activity in PeerWise typically increases rapidly before a test (Denny, 2015), and that answering questions is strongly predictive of subsequent test performance (Denny et al., 2018, Snow et al., 2019). Instructors are also able to make use of the questions, for example by reviewing a question repository to identify topics that are challenging for students, or by selecting high-quality questions for use on summative tests and exams. For example, Huang et al. (2021) showed that with some basic coaching, students using PeerWise were able to produce many questions that performed just as well when used on high-stakes exams as questions authored by academics.

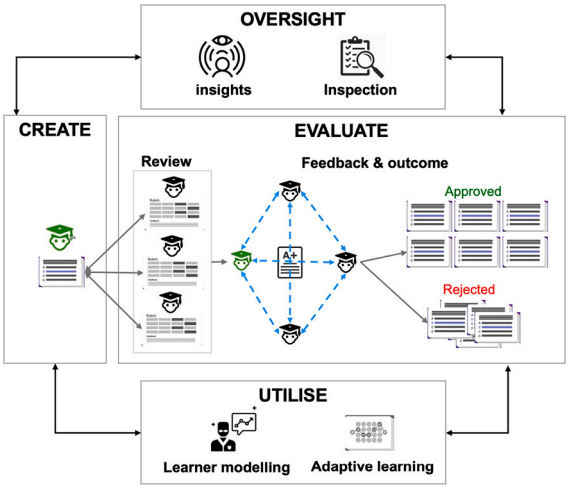


Fig. 7. Overview of RiPPLE.

3.1.4. Oversight in PeerWise

PeerWise was originally inspired by the contributing student approach described by (Hamer, 2006), and thus takes the view that students are primarily responsible for both producing and moderating the resources. As such, instructor oversight of the learnersourcing process is fairly minimal.

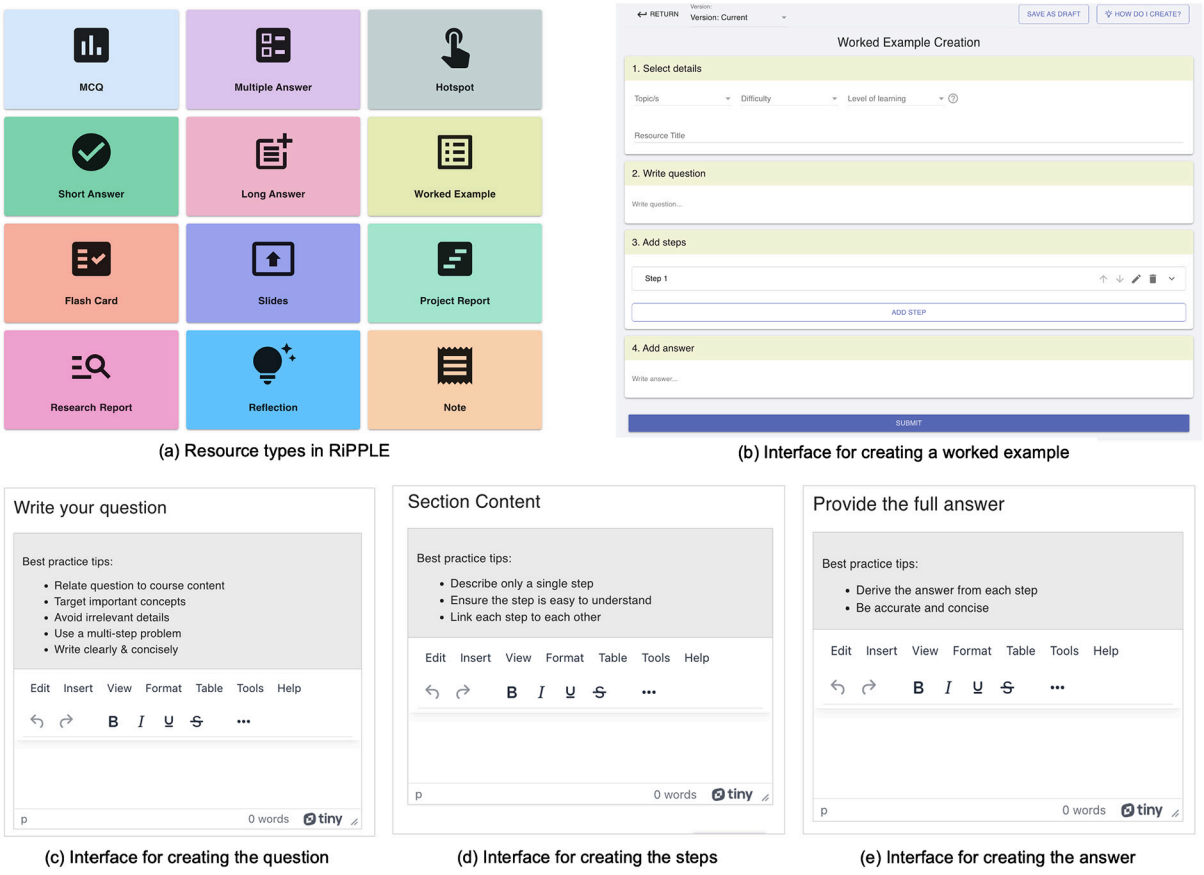
To provide some initial structure for students when authoring questions, instructors can define a set of "course tags" which are topics that are shown to students during the authoring step at the point they are prompted to tag their question. Defining these tags can help to minimise fragmentation of topics. Instructors can also post "administrator comments" as part of the comment thread for any question. Such comments are highlighted as being posted by the instructor, and appear separately from student written comments.

Students can, of course, edit their own questions but cannot edit questions written by other students. They are limited to providing feedback via a question's comment thread or flagging questions, at the time of rating, that they deem are inappropriate. Instructors have the ability to edit or delete questions in the repository, and can review questions which have low quality ratings or which have been flagged by students.

3.2. RiPPLE

A full description of RiPPLE is provided in Khosravi et al. (2019). Here, we provide a brief description based on our proposed learnersourcing framework presented in Fig. 1. At its core, RiPPLE is an adaptive educational system that relies on learnersourcing for content creation. RiPPLE can be used as a standalone system or be integrated into many popular LMSs using the Learning Tools Interoperability (LTI) standard. RiPPLE supports two types of roles: instructors and students. To use the system in a course, an instructor creates a RiPPLE offering and adds a set of topics, and optionally imports resources from other RiPPLE offerings. This enables instructors to import resources from their past offerings as well as share resources with other instructors with or outside their institution. In RiPPLE, students own the intellectual property rights of any content that they create and are free to share their content with others as they desire. However, to enable students to benefit from each other's contributions, the terms and conditions of using the platform request students to provide a non-exclusive licence to host, use, distribute, modify, run, copy and publicly display their content.

Since 2018, RiPPLE has been used in over 150 course offerings with over 30k students who have created over 80k resources which have received over 300k peer reviews. Fig. 7 provides an overview of the



(a) Resource types in RiPPLE

(b) Interface for creating a worked example

(c) Interface for creating the question

(d) Interface for creating the steps

(e) Interface for creating the answer

Fig. 8. Creation in RiPPLE.

operation of RiPPLE with respect to our framework. We elaborate on each part of the framework in the following sections.

3.2.1. *Create in RiPPLE*

Both students and instructors can create learning resources in RiPPLE. Originally, similar to many other learnersourcing systems, RiPPLE only supported the creation of MCQs. However, during the last few years, as demonstrated by Fig. 8(a), we have added support for the creation of various other educational resource types such as flashcards, hotspots, worked examples as well as short and long answer questions. Based on popular demand from instructors, we have also included resource types such as reflections, research reports and project reports, which allow students to share original contributions that are not directly related to covering content from a course curriculum. Fig. 8(b) illustrates the interface used for creating a worked example. Students are asked to identify the topic(s) associated with the resource, its level of difficulty, and it's corresponding position within Bloom's revised taxonomy of learning objectives (Krathwohl, 2002). Given that we expect not all students to be familiar with this taxonomy, the platform provides a description and an example resource for each level of learning. Fig. 8(c)-(e) show the interfaces and the provided best practice tips for writing the question, steps and a sample solution.

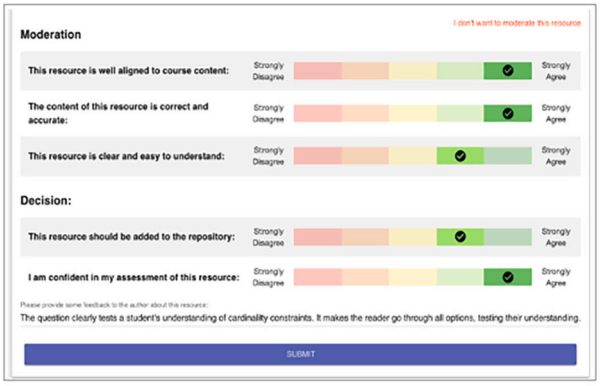
3.2.2. *Evaluate in RiPPLE*

Resources created by students go through a formal peer-review process (Darvishi et al., 2021). Upon availability of a student to peer review a resource (i.e., the student goes on the moderation tab on the platform), RiPPLE selects and presents a non-evaluated resource to the student. Each resource type has an associated rubric for evaluation, and these rubrics share a similar structure. All rubrics have a set of criteria, a statement capturing the evaluators' perceptions of the overall quality of the resource, a statement capturing their confidence in their judgement and written feedback to justify their decisions. However, specific details the underlying criteria for evaluating various resource types have changed over time. The initial version of the rubric, shown in Fig. 9(a), used a Likert scale to capture students' responses to criteria (alignment with course content, correctness and coherence of the resource), decision and confidence level. Analysis of more than 40,000 student evaluations based on this rubric revealed that it led students to lenient marking as over half of the students provided the highest quality rating to the resources.[4]

We also noticed that the average length of the provided comments was only eight words, which meant very little support was provided for their judgement. To address some of these shortcomings, we updated the rubric to what is shown in Fig. 9(b). This rubric included additional criteria that referred to the appropriate level of difficulty and critical thinking. In addition, we moved away from Likert scale statements, which are commonly used to capture perceptions in surveys, to words that refer to the quality of outcome ranging from poor to outstanding, which is more commonly used in rubrics. Finally, the updated rubric specifically asked students to justify their decision and provide feedback rather than just having space for a comment without specific instructions. Analysis of more than 30,000 student evaluations based on the updated rubric showed a significant shift in their responses where the most common rating response moved from the highest rating (5) to the second highest rating (4). The mean length of the associated comments supporting the quality ratings also increased to 13 words.

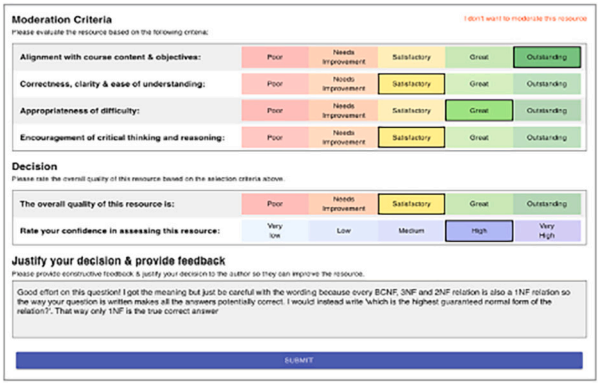
Fig. 9(c) illustrates the next main update to the rubric, described in (Darvishi, Khosravi, Abdi, et al., 2022), which aimed to improve

[4] Informed consent from students and approval from The University of Queensland Human Research Ethics Committee was received for reporting aggregated stats on human participants' engagement.



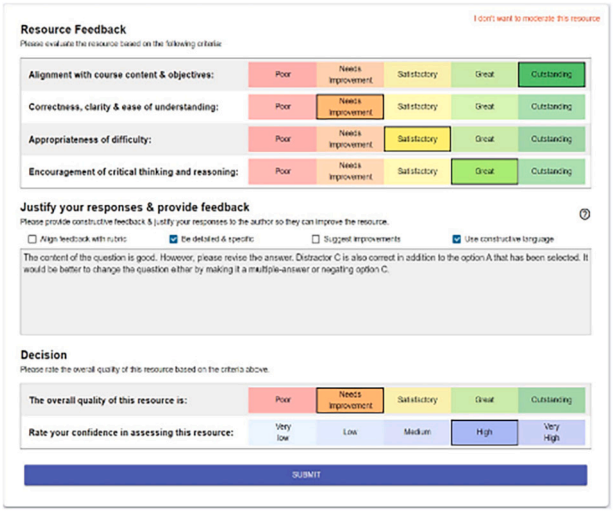
# responses	Decision rating					Avg comment length
	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	
41,048	1%	3%	9%	35%	51%	8

(a) Initial rubric with likert scale



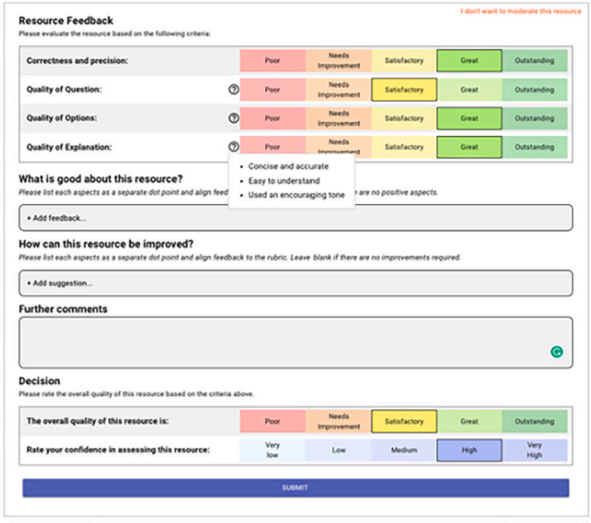
# responses	Decision rating					Avg comment length
	Poor	Needs improvement	Satisfactory	Great	Outstanding	
31,012	1%	3%	23%	58%	15%	13

(b) Interface redesign with more room for feedback and move to quality criteria



# responses	Decision rating					Avg comment length
	Poor	Needs improvement	Satisfactory	Great	Outstanding	
191,928	1	5	23	55	17	25

(c) Addition of more detailed criteria, feedback tips and automated control prompt



# responses	Decision rating					Avg comment length
	Poor	Needs improvement	Satisfactory	Great	Outstanding	
4,026	1	4	18	51	26	31

(d) Scaffolding feedback to positive feedback plus suggestions for improvement

Fig. 9. Evaluate in RiPPLE.

the quality of the provided feedback. Informed by higher education research, we built a set of training materials (accessible by clicking on the? button next to where they justify their response) and a self-monitoring checklist for students to consider while writing their reviews. We also developed natural language processing-based quality control functions that automatically assess feedback submitted and prompt students to improve, if necessary. Analysis of over 190,000 student evaluations based on this rubric indicated that we were successful in almost doubling the length of the provided feedback from an average of 13 to 25 words.

Fig. 9 (d) shows the current version of the rubric on the platform which has recently been deployed. This incorporates two new changes. The first change is that each criterion can now be accompanied by a list of items for the reviewer to consider in their evaluation. This change allows the platform or instructors to provide more elaborate guidelines for a reviewer's consideration. The second change has introduced scaffolding to the feedback part where instead of writing one block of feedback students now provide a list of positive aspects about the resources followed by a list of suggestions. They can then provide further comments, if necessary. This change was introduced as the feedback from many reviewers was generic in nature and did not include constructive suggestions on how the resource can be improved. Analysis of the impact of this new rubric is underway. Early results based on over 4,000 responses are included in Fig. 9 (d).

A feedback and evaluation outcome interface (as shown in Fig. 10), shares the results with the author and the reviewers, asking them to vote on the helpfulness of the evaluations and determine whether or not they agree with the outcome (approved or denied). We have collected ore than 30,000 responses since the interface was added to the platform. It is encouraging to see that students generally trust the system; only 2% of the responses disagreed with the outcomes of the peer assessment process and fewer than 4% mentioned they were unsure (see the work of Darvishi, Khosravi, Sadiq, et al. (2022) for more details).

3.2.3. *Utilise in RiPPLE*

Fig. 11 illustrates the interface used for providing personalised practice opportunities for students. The top part of the figure represents an interactive visualisation widget, in the form of an open learner model (Bull, 2020, Abdi, Khosravi, Sadiq, Darvishi, 2021), that allows students to view an abstract representation of their knowledge state based

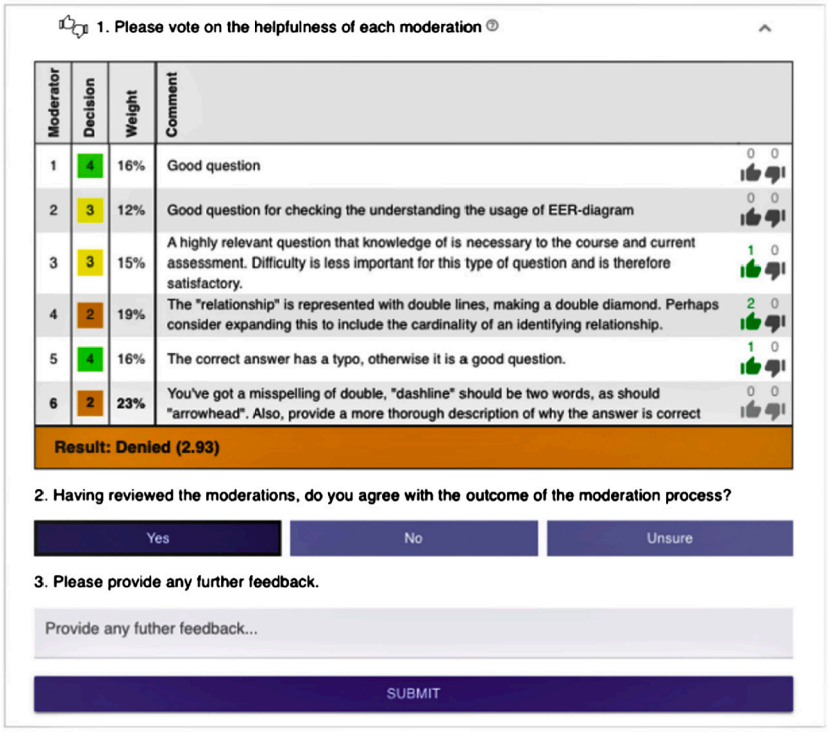


Fig. 10. Feedback in RiPPLE.

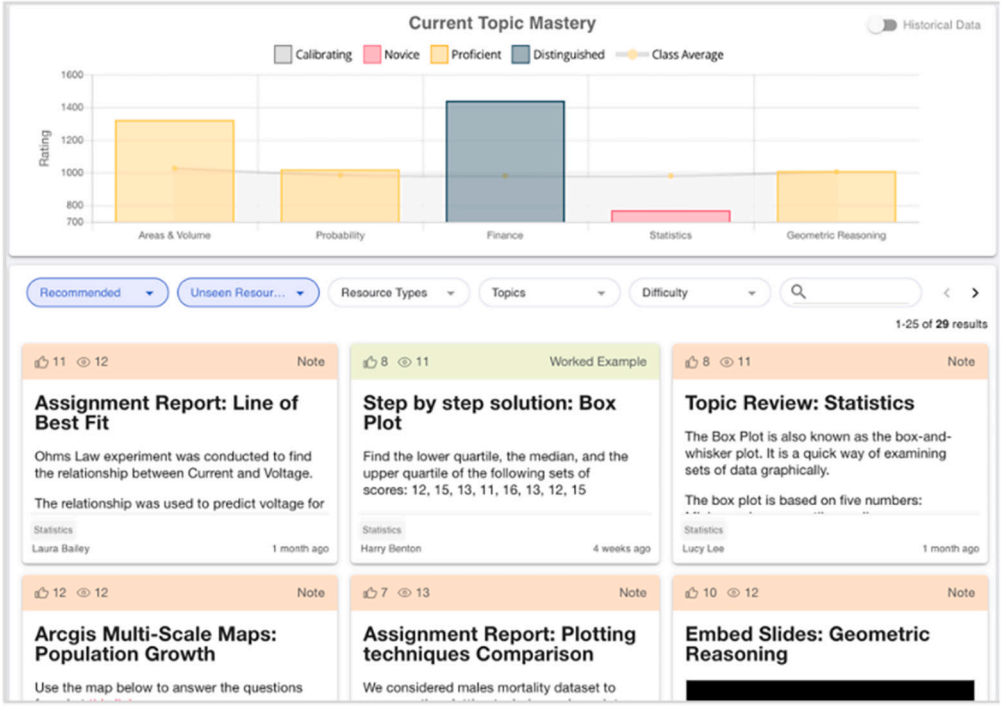


Fig. 11. The open learner model and recommendation interface in RiPPLE.

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. Namely, for a particular unit of knowledge, red, yellow and green signify inadequate competence, adequate competence with room for improvement, and mastery, respectively. Currently, RiPPLE employs an Elo-based rating system for approximating the knowledge state of users with the results translated into coloured bars (Abdi, Khosravi, Sadiq, 2020). The lower part of the screen displays learning content from the repository of approved resources that are recommended to a student based on their learning needs using the recommender system outlined in (Khosravi et al., 2017). At a high level, this system recommends easier content on topics where students are developing mastery and harder content on topics for which students have already developed mastery. Students also have the ability to search for resources based on various criteria such as the resource type, topics and difficulty.

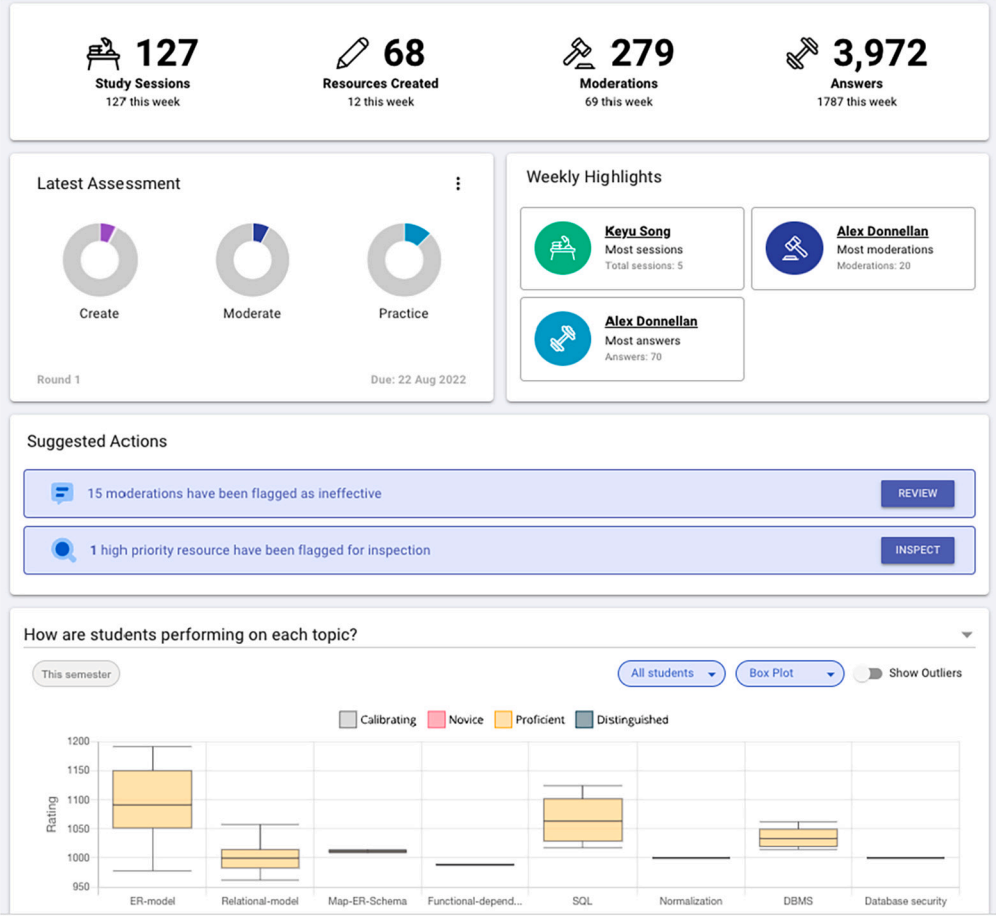


Fig. 12. Weekly insights in RiPPLE.

3.2.4. *Oversight in RiPPLE*

One of the main design guidelines of RiPPLE is to ensure that it optimally uses the minimal availability of instructors. To do so, RiPPLE has an instructor landing page, shown in Fig. 12. The top part of the page displays statistics based on overall use and use from the previous week on study sessions as well as resources created, evaluated and answered. It then provides information about the completion status for the latest round of assessment and weekly highlights about students' achievements and popular resources. One of the main ways we have tried to optimally use instructor time is to add suggested actions. RiPPLE currently provides five types of suggested actions including inspecting resources that benefit the most from academic judgement, reviewing evaluations that are flagged as ineffective, submitting assessment grades, nudging at-risk students (e.g., those who have not logged in or completed assessment) and congratulating high achievers on their achievements. The bottom of the page includes an analytical toolbox that provides answers to a list of questions in relation to students' performance and engagement. For each of the questions instructors can set the start and end date for data being reported to view class-level or individual-level trends using various visualisation types (e.g., bar charts, box plots).

Fig. 13 shows the underlying interface for two of the main suggested actions, namely inspecting resources and reviewing evaluations. Fig. 13(a) shows the RiPPLE interface for inspecting resources that have been approved but are likely to be incorrect or ineffective. At a high level, it employs a range of human-driven metrics (e.g., high disagreement in reviews, a high ratio of downvotes in comparison to upvotes) and data-driven metrics (e.g., assessment items that have a low discrimination index or questionable distractors where the popular answer is not the one proposed by the author) to categorise resources into having high, medium, low or no priority for being reviewed. RiPPLE uses absolute and relative points of comparison to help instructors make sense of why a resource has been flagged for review (e.g., "disagreements between student ratings for this resource are 2.8 times higher than average of the course").

Fig. 13(b) shows the RiPPLE interface for searching the reviews. It enables instructors to set a date range to identify reviews with a particular word count range on a particular set of topics. In addition, we use the functions outlined in (Darvishi, Khosravi, Abdi, et al., 2022) to enable instructors to identify whether or not reviews include a suggestion. Once instructors select a set of reviews, they can apply one of the following actions: upvote it to provide positive feedback; ignore it so that it wouldn't show up in the search again; downvote it and provide feedback on its ineffectiveness; or finally remove the review so that it wouldn't count towards meeting assessment requirements for the reviewer.

4. Challenges and implications for practice and research

The adoption of our learnersourcing framework may lead to certain considerations that need to be addressed with care. This section elaborates on these challenges and implications, and also pinpoints areas that call for further exploration and research.

The process of educating students to become proficient creators and evaluators of content, along with the duty of content creation itself, are tasks that demand considerable time. Therefore educators that are considering taking on learnersourcing may need to think about how these responsibilities can be balanced against other academic commitments in their course to avoid overloading students. Moreover, some students may not fully understand or appreciate the advantages of learnersourc-

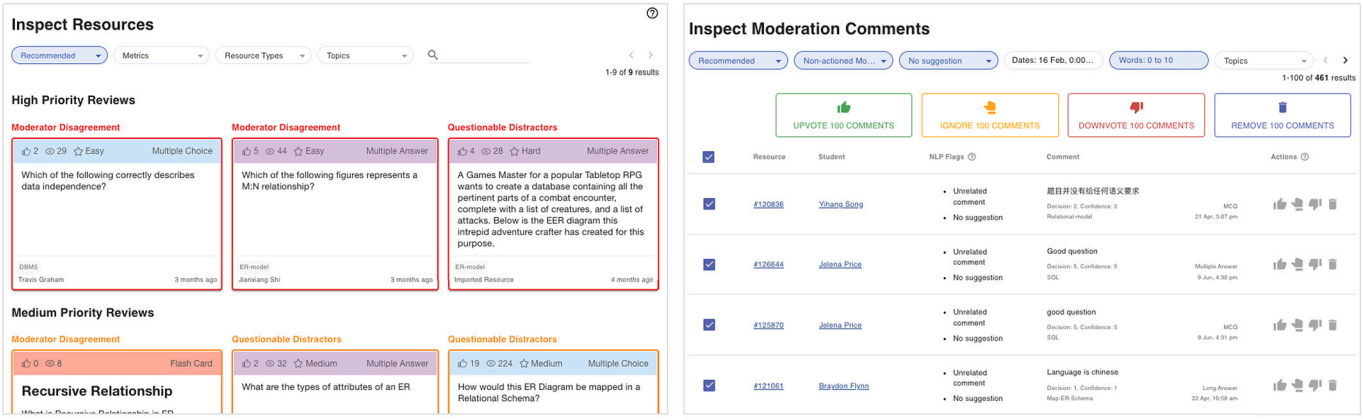


Fig. 13. Resource and peer review inspection in RiPPLE. The name of students has redacted in the figure.

ing. Thus, incentivising them to participate may necessitate educators to explicitly discuss the benefits and rationale for its inclusion in their curriculum. Empirical studies focusing on the most effective strategies for incetivising students and training them in content creation, as well as assessing the impact of these strategies on student engagement, outcomes, and the quality of resources produced, could provide substantial insights for effective adoption of learnersourcing.

The utilisation of AI assistance to aid students in content creation and evaluation is a potentially beneficial and promising strategy. However, if employed without care, it may lead to overdependence, where students might excessively rely on large language models to create and evaluate content on their behalf. This could result in several negative outcomes, such as content lacking originality, not aligning with course material, or simply being factually incorrect. The design, development and validation of interfaces and scaffolds incorporating large language model APIs into learnersourcing systems, to help students more effectively contribute, presents a promising field of future research.

While the frameworks include multiple strategies to optimise the utilisation of the instructor's efforts, a certain level of monitoring is still required to identify misbehaviour such as carrying out detrimental peer reviews, plagiarising content, or creating inappropriate resources. This monitoring not only serves as a means of maintaining content and behavioural standards, but also serves as a demonstration of reliability, providing assurance to both educators and students that the system is trustworthy and dependable. In more minor cases, the created content may lack relevance or have large coverage gaps based on the course objectives. These actions could undermine students' trust in learnersourcing and seeing its benefits. The design, development and validation of strategies that provide actionable insights and pedagogical interventions that best assist educators in facilitating learning, identifying students who require assistance and overseeing the creation and evaluation of contributions of students in learnersourcing systems, all while demonstrating their reliability, hold substantial potential for future research

Utilising models that foster content co-creation among several students might provide the dual advantages of promoting collaborative learning and enhancing the quality of content produced. However, this approach is also prone to issues such as group dynamic conflicts, free-riding, ensuring fairness in grading, and evaluating individual contributions which are inherent to group-based learning and assessment tasks. Educators may need to think about approaches for dealing with such conflicts if they are considering enabling multi-student learnersourced submissions.

5. Conclusion

Widespread changes to learning and teaching alongside rapid growth in the use of digital tools in education and the vast data they collect are presenting new opportunities for the application of artificial intelligence in the classroom. One activity that appears uniquely placed to benefit is learnersourcing. Given that it is primarily student-driven, artificial models of intelligence can be employed to enable students to develop transformative competencies, such as creating new value, developing self- and co-agency skills and improving the learning experience while reducing the need for instructor expertise and oversight.

This paper introduced a framework that considers the existing learnersourcing literature, the latest insights from the learning sciences and advances in AI to offer a blueprint for developing learnersourcing systems. The framework is presented in the form of questions that correspond to creating content, evaluating the quality of the created content, considerations for how learnersourced contributions can be utilised and the role of instructors in facilitating learning via learnersourcing.

A common theme across all four dimensions is the need for human-AI partnerships for advancing learnersourcing. In terms of content creation and evaluation, advances in NLP and generative models provide space for AI to play a fundamental role in the co-creation of content with humans and to assist with the automated evaluation of its quality. For utilising learnersourcing content, the use of AI can help in developing explainable recommender systems that model students' mastery and assist them in engaging with content that best suits their learning needs. Finally, which respect to oversight, AI can partner with instructors to help them optimally use their time in reviewing content that benefits the most from their judgement and assisting students that need their help. To maintain a clear connection between the discussion and the literature pertaining to each dimension of the framework, we have included discussion points within the sections where they are presented. The two presented case studies demonstrate the application of our framework in the context of two vastly different and well-adopted learnersourcing systems.

Despite a long history of exploration, the development and large-scale adoption of learnersourcing systems are still in their early stages and much more fundamental work is needed before they can achieve their full potential. We call upon the educational research and practitioner communities to review and critique our framework and to contribute to advancing the field through the development of scientifically grounded and empirically validated systems that can help in accelerating the development and adoption of learnersourcing. In relation to human-AI partnership for creating and evaluating content, we highlight many opportunities, but there is still much that needs to be done for this vision to reach its full potential. On the educational side, there is a need for preparing students and instructors for collaboration with AI, which requires large-scale upskilling and training programs in data and digital literacy. There is also a need for conducting educational research to determine the effect size of various human-AI partnerships in creating content on student learning. Relative to social science, there is a need for developing new policies that assign ownership of content

and copyright among human and AI collaborators in a fair manner. On the technical side, more effective methods of assisting humans in the create, evaluate, utilise and oversight processes are required. In particular, development of discipline-focused large language models may strengthen the quality of novel educational content AI can create. Finally, in relation to ethics, there is a need for exploring spot-checking and human-in-the-loop models that fairly treat all students and gives them equal opportunity for learning and receiving feedback.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdi, S. (2022). Learner models for learnersourced adaptive educational systems. Ph.D. thesis, The University of Queensland.

Abdi, S., Khosravi, H., & Sadiq, S. (2020). Modelling learners in crowdsourcing educational systems. In *International conference on artificial intelligence in education* (pp. 3–9). Springer.

Abdi, S., Khosravi, H., Sadiq, S., & Darvishi, A. (2021). Open learner models for multi-activity educational systems. In *International conference on artificial intelligence in education* (pp. 11–17). Springer.

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

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate elo-based learner model for adaptive educational systems. In *Proc. educational data mining conf.* (pp. 462–467).

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Complementing educational recommender systems with open learner models. In *Proc. 10th int. conf. learning analytics knowledge* (pp. 360–365).

Ajjawi, R., & Boud, D. (2017). Researching feedback dialogue: An interactional analysis approach. *Assessment & Evaluation in Higher Education*, *42*, 252–265. https://doi.org/10.1080/02602938.2015.1102863.

Aleven, V., McLaren, B. M., Sewall, J., Van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education*, *26*, 224–269. https://doi.org/10.1007/s40593-015-0088-2.

Antin, J. (2011). My kind of people? Perceptions about Wikipedia contributors and their motivations. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3411–3420).

Bada, S. O., & Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education*, *5*, 66–70. https://doi.org/10.29322/IJSRP.12.12.2022.p13211.

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

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., & Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on user interface software and technology* (pp. 313–322).

Bhatnagar, S., Zouaq, A., Desmarais, M. C., & Charles, E. (2020). Learnersourcing quality assessment of explanations for peer instruction. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S. M. Dennerlein (Eds.), *Addressing global challenges and quality education* (pp. 144–157). Cham: Springer International Publishing.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*, 703–725. https://doi.org/10.1007/s10648-018-9434-x.

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models, https://doi.org/10.48550/ARXIV.2108.07258.

Bredow, C. A., Roehling, P. V., Knorp, A. J., & Sweet, A. M. (2021). To flip or not to flip? A meta-analysis of the efficacy of flipped learning in higher education. *Review of Educational Research*, *91*, 878–918. https://doi.org/10.3102/00346543211019122.

Bull, S. (2020). There are open learner models about! *IEEE Transactions on Learning Technologies*, *13*, 425–448. https://doi.org/10.1109/TLT.2020.2978473.

Cambre, J., Klemmer, S., & Kulkarni, C. (2018). Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13).

Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education*, *34*, 79–89. https://doi.org/10.1080/02602930801895786.

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment & Evaluation in Higher Education*, *44*, 705–714. https://doi.org/10.1080/02602938.2018.1531108.

Carless, D., Chan, K. K. H., To, J., Lo, M., & Barrett, E. (2018). Developing students' capacities for evaluative judgement through analysing exemplars. In *Developing evaluative judgement in higher education* (pp. 108–116). Routledge.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In *Learning in doing: Social, cognitive and computational perspectives* (pp. 39–64). Cambridge University Press.

Chan, A. (2023). Gpt-3 and instructgpt: Technological dystopianism, utopianism, and "contextual" perspectives in ai ethics and industry. *AI and Ethics*, *3*, 53–64. https://doi.org/10.1007/s43681-022-00148-6.

Chang, J. C., Amershi, S., & Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 2334–2346). New York, NY, USA: ACM.

Chao, M. M., Visaria, S., Mukhopadhyay, A., & Dehejia, R. (2017). Do rewards reinforce the growth mindset?: Joint effects of the growth mindset and incentive schemes in a field intervention. *Journal of Experimental Psychology. General*, *146*, 1402. https://doi.org/10.1037/xge0000355.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. https://doi.org/10.48550/arXiv.2107.03374.

Choi, K., Shin, H., Xia, M., & Kim, J. (2022). Algosolve: Supporting subgoal learning in algorithmic problem-solving with learnersourced microtasks. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. New York, NY, USA: ACM.

Collis, B., & Moonen, J. (2002). The contributing student: A pedagogy for flexible learning. *Computers in the Schools*, *19*, 207–220. https://doi.org/10.1300/J025v19v03_16.

Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*, 669. https://doi.org/10.1037/0278-7393.15.4.669.

Darvishi, A., Khosravi, H., Abdi, S., Sadiq, S., & Gašević, D. (2022). Incorporating training, self-monitoring and ai-assistance to improve peer feedback quality. In *Proceedings of the ninth ACM conference on learning @ scale* (pp. 35–47).

Darvishi, A., Khosravi, H., Rahimi, A., Sadiq, S., & Gašević, D. (2023). Assessing the quality of student-generated content at scale: A comparative analysis of peer-review models. *IEEE Transactions on Learning Technologies*, *16*, 106–120. https://doi.org/10.1109/TLT.2022.3229022.

Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the eighth ACM conference on learning @ scale* (pp. 139–150).

Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13233.

Denny, P. (2013). The effect of virtual achievements on student engagement. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 763–772). New York, NY, USA: ACM.

Denny, P. (2015). Generating practice questions as a preparation strategy for introductory programming exams. In *Proceedings of the 46th ACM technical symposium on computer science education* (pp. 278–283). New York, NY, USA: ACM.

Denny, P., Cukierman, D., Luxton-Reilly, A., & Tempero, E. (2012). A case study of multi-institutional contributing-student pedagogy. *Computer Science Education*, *22*, 389–411. https://doi.org/10.1080/08993408.2012.727712.

Denny, P., Luxton-Reilly, A., & Hamer, J. (2008). The peerwise system of student contributed assessment questions. In *Proceedings of the tenth conference on Australasian computing education, vol. 78* (pp. 69–74). AUS: Australian Computer Society, Inc.

Denny, P., Luxton-Reilly, A., & Simon, B. (2009). Quality of student contributed questions using peerwise. In *Proceedings of the eleventh Australasian conference on computing education, vol. 95* (pp. 55–63).

Denny, P., Luxton-Reilly, A., Tempero, E., & Hendrickx, J. (2011). Codewrite: Supporting student-driven practice of Java. In *Proceedings of the 42nd ACM technical symposium on computer science education* (pp. 471–476). New York, NY, USA: ACM.

Denny, P., McDonald, F., Empson, R., Kelly, P., & Petersen, A. (2018). Empirical support for a causal relationship between gamification and learning outcomes. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery.

Denny, P., Sarsa, S., Hellas, A., & Leinonen, J. (2022). Robosourcing educational resources–leveraging large language models for learnersourcing. In *Proceedings of the first annual workshop on learnersourcing: Student-generated content @ scale*.

Devine, D. J., & Kozlowski, S. W. (1995). Domain-specific knowledge and task characteristics in decision making. *Organizational Behavior and Human Decision Processes*, *64*, 294–306.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*, 945–955. https://doi.org/10.3758/BF03196872.

Divate, M., & Salgaonkar, A. (2017). Automatic question generation approaches and evaluation techniques. *Current Science*, 1683–1691. https://doi.org/10.18520/cs/v113/i09/1683-1691.

Doroudi, S., Kamar, E., Brunskill, E., & Horvitz, E. (2016). Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2623–2634).

Doroudi, S., Williams, J. J., Kim, J., Patikorn, T., Ostrow, K., Selent, D., Heffernan, N. T., Hills, T. T., & Rosé, C. P. (2018). Crowdsourcing and education: Towards a theory and praxis of learnersourcing. In M. Mavrikis, & K. Porayska-Pomsta (Eds.), *Rethinking learning in the digital age: Making the learning sciences count - proceedings of the 13th international conference of the learning sciences, ICLS 2018*. International Society of the Learning Sciences. https://repository.isls.org/handle/1/603.

Doyle, E., Buckley, P., & Whelan, J. (2019). Assessment co-creation: An exploratory analysis of opportunities and challenges based on student and instructor perspectives. *Teaching in Higher Education*, *24*, 739–754. https://doi.org/10.1080/13562517.2018.1498077.

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, et al. (2021). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level, https://doi.org/10.48550/ARXIV.2112.15594.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology, vol. 44* (pp. 247–296). Elsevier.

Galloway, K. W., & Burns, S. (2015). Doing it for themselves: Students creating a high quality peer-learning environment. *Chemistry Education Research and Practice*, *16*, 82–92. https://doi.org/10.1039/C4RP00209A.

Gao, X. A., Wright, J. R., & Leyton-Brown, K. (2019). Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence*, *275*, 618–638. https://doi.org/10.1016/j.artint.2019.03.004.

Gehringer, E. F., Ehresman, L. M., & Skrien, D. J. (2006). Expertiza: Students helping to write an OOD text. In *Companion to the 21st ACM SIGPLAN symposium on object-oriented programming systems, languages, and applications* (pp. 901–906). New York, NY, USA: Association for Computing Machinery.

Geiger, D., Rosemann, M., & Fielt, E. (2011). Crowdsourcing information systems–a systems theory perspective. In *22nd Australasian conference on information systems*.

Glassman, E. L., Lin, A., Cai, C. J., & Miller, R. C. (2016). Learnersourcing personalized hints. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 1626–1636).

Guo, P. J., Markel, J. M., & Zhang, X. (2020). Learnersourcing at scale to overcome expert blind spots for introductory programming: A three-year deployment study on the python tutor website. In *Proceedings of the seventh ACM conference on learning @ scale* (pp. 301–304).

Gyamfi, G., Hanna, B., & Khosravi, H. (2021a). Supporting peer evaluation of student-generated content: A study of three approaches. *Assessment & Evaluation in Higher Education*, 1–19. https://doi.org/10.1080/02602938.2021.2006140.

Gyamfi, G., Hanna, B. E., & Khosravi, H. (2021b). The effects of rubrics on evaluative judgement: A randomised controlled experiment. *Assessment & Evaluation in Higher Education*, *47*, 126–143. https://doi.org/10.1080/02602938.2021.1887081.

Hamer, J. (2006). Some experiences with the "contributing student approach". In *Proceedings of the 11th annual SIGCSE conference on innovation and technology in computer science education* (pp. 68–72). New York, NY, USA: ACM.

Hamer, J., Cutts, Q., Jackova, J., Luxton-Reilly, A., McCartney, R., Purchase, H., Riedesel, C., Saeli, M., Sanders, K., & Sheard, J. (2008). Contributing student pedagogy. *SIGCSE Bulletin*, *40*, 194–212. https://doi.org/10.1145/1473195.1473242.

Hamer, J., Ma, K. T., & Kwong, H. H. (2005). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian conference on computing education, vol. 42* (pp. 67–72).

Hamer, J., Sheard, J., Purchase, H., & Luxton-Reilly, A. (2012). Contributing student pedagogy, https://doi.org/10.1145/1473195.1473242.

Hardy, J., Bates, S. P., Casey, M. M., Galloway, K. W., Galloway, R. K., Kay, A. E., Kirsop, P., & McQueen, H. A. (2014). Student-generated content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*, *36*, 2180–2194. https://doi.org/10.1080/09500693.2014.916831.

Heffernan, N. T., & Heffernan, C. L. (2014). The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*, 470–497. https://doi.org/10.1007/s40593-014-0024-x.

Hills, T. T. (2015). Crowdsourcing content creation in the classroom. *Journal of Computing in Higher Education*, *27*, 47–67. https://doi.org/10.1007/s12528-015-9089-2.

Hilton, C. B., Goldwater, M. B., Hancock, D., Clemson, M., Huang, A., & Denyer, G. (2022). Scalable science education via online cooperative questioning. *CBE—Life Sciences Education*, *21*, Article ar4. https://doi.org/10.1187/cbe.19-11-0249.

Huang, A., Hancock, D., Clemson, M., Yeo, G., Harney, D., Denny, P., & Denyer, G. (2021). Selecting student-authored questions for summative assessments. *Research in Learning Technology*, *29*, 1–25. https://doi.org/10.25304/rlt.v29.2517.

Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of cscl in stem education during 2005–2014. *Educational Research Review*, *28*, Article 100284. https://doi.org/10.1016/j.edurev.2019.100284.

Ji, T., Lyu, C., Jones, G., Zhou, L., & Graham, Y. (2022). Qascore—an unsupervised unreferenced metric for the question generation evaluation. *Entropy*, *24*, 1514. https://doi.org/10.3390/e24111514.

Jiang, Y., Schlagwein, D., & Benatallah, B. (2018). A review on crowdsourcing for education: State of the art of literature and practice. In *PACIS* (p. 180).

Jin, H., Chang, M., & Kim, J. (2019). Solvedeep: A system for supporting subgoal learning in online math problem solving. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1–6).

Jin, H., & Kim, J. (2022). Learnersourcing subgoal hierarchies of code examples. In *Proceedings of the first annual workshop on learnersourcing: Student-generated content @ scale*.

Jørnø, R. L., & Gynther, K. (2018). What constitutes an 'actionable insight'in learning analytics? *Journal of Learning Analytics*, *5*, 198–221. https://doi.org/10.18608/jla.2018.53.13.

Kaliisa, R., Rienties, B., Mørch, A. I., & Kluge, A. (2022). Social learning analytics in computer-supported collaborative learning environments: A systematic review of empirical studies. *Computers and Education Open*, Article 100073. https://doi.org/10.1016/j.caeo.2022.100073.

Kao, G. Y. M. (2013). Enhancing the quality of peer review by reducing student "free riding": Peer assessment with positive interdependence. *British Journal of Educational Technology*, *44*, 112–124. https://doi.org/10.1111/j.1467-8535.2011.01278.x.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kay, A. E., Hardy, J., & Galloway, R. K. (2020). Student use of peerwise: A multi-institutional, multidisciplinary evaluation. *British Journal of Educational Technology*, *51*, 23–35. https://doi.org/10.1111/bjet.12754.

Kelley, M. R., Chapman-Orr, E. K., Calkins, S., & Lemke, R. J. (2019). Generation and retrieval practice effects in the classroom using peerwise. *Teaching of Psychology*, *46*, 121–126. https://doi.org/10.1177/0098628319834174.

Khan, V. J., Papangelis, K., & Markopoulos, P. (2020). Completing a crowdsourcing task instead of an assignment; what do university students think? In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–8).

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

Khosravi, H., Demartini, G., Sadiq, S., & Gasevic, D. (2021). Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 32–42).

Khosravi, H., Gyamfi, G., Hanna, B. E., Lodge, J., & Abdi, S. (2021). Bridging the gap between theory and empirical research in evaluative judgment. *Journal of Learning Analytics*, *8*, 117–132. https://doi.org/10.18608/jla.2021.7206.

Khosravi, H., Kitto, K., & Williams, J. J. (2019). Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, *6*, 91–105. https://doi.org/10.18608/jla.2019.63.12.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074.

Kim, H., Song, I., & Kim, J. (2022). Learnersourcing modular and dynamic multiple choice questions. In *Learnersourcing: Student-generated content @ scale*. Springer.

Kim, J. (2015). Learnersourcing: Improving learning with collective learner activity. Ph.D. thesis, Massachusetts Institute of Technology.

Kim, J., Guo, P. J., Cai, C. J., Li, S. W., Gajos, K. Z., & Miller, R. C. (2014). Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on user interface software and technology* (pp. 563–572).

Kim, J., Miller, R. C., & Gajos, K. Z. (2013). Learnersourcing subgoal labeling to support learning from how-to videos. In *CHI 2013* (pp. 685–690). New York, NY, USA: ACM.

Kinjo, H., & Snodgrass, J. G. (2000). Does the generation effect occur for pictures? *The American Journal of Psychology*, *113*, 95. https://doi.org/10.2307/1423462.

Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, *34*, 27–41. https://doi.org/10.1609/aimag.v34i3.2484.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*, 212–218. https://doi.org/10.1207/s15430421tip4104_2.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y.

Lahza, H., Khosravi, H., & Demartini, G. (2023). Analytics of learning tactics and strategies in an online learnersourcing environment. *Journal of Computer Assisted Learning*, *39*, 94–112. https://doi.org/10.1111/jcal.12729.

Lahza, H., Khosravi, H., Demartini, G., & Gasevic, D. (2022). Effects of technological interventions for self-regulation: A control experiment in learnersourcing. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 542–548).

Lee, C., et al. (2020). Question generation workflow: Incorporating student-generated content and peer evaluation. Ph.D. thesis, Massachusetts Institute of Technology.

Lehtinen, E., Hakkarainen, K., Lipponen, L., Rahikainen, M., & Muukkonen, H. (1999). Computer supported collaborative learning: A review. In *The JHGI Giesbers reports on education, vol. 10*.

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., Tran, A., & Hellas, A. (2023). *Comparing code explanations created by students and large language models*In *Proceedings of the 28th ACM conference on innovation and technology in computer science education, vol. 1*. New York, NY, USA: ACM.

Leinonen, J., Pirttinen, N., & Hellas, A. (2020). Crowdsourcing content creation for sql practice. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education* (pp. 349–355). New York, NY, USA: Association for Computing Machinery.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*, 525–536. https://doi.org/10.1111/j.1467-8535.2009.00968.x.

Liu, Y., & Chen, Y. (2016). Learning to incentivize: Eliciting effort via output agreement. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 3782–3788).

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*, 30–43. https://doi.org/10.1016/j.jslw.2008.06.002.

Malau-Aduli, B. S., Assenheimer, D., Choi-Lundberg, D., & Zimitat, C. (2014). Using computer-based technology to improve feedback to staff and students on mcq assessments. *Innovations in Education and Teaching International*, *51*, 510–522. https://doi.org/10.1080/14703297.2013.796711.

Matcha, W., Gašević, D., Jovanović, J., Uzir, N. A., Oliver, C. W., Murray, A., & Gasevic, D. (2020). Analytics of learning strategies: The association with the personality traits. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 151–160).

Matcha, W., Gašević, D., Pardo, A., et al. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*, 226–245. https://doi.org/10.1109/TLT.2019.2916802.

McBroom, J., & Paassen, B. (2020). Assessing the quality of mathematics questions using student confidence scores. In J. Wang, A. Lamb, E. Saveliev, P. Cameron, et al. (Eds.), *Winning contributions for task 3 of the NeurIPS 2020 education challenge*. https://dqanonymousdata.blob.core.windows.net/neurips-public/papers/mcbroom-paassen/neurips_edu_2020.pdf.

Mitros, P. (2015). Learnersourcing of complex assessments. In *L@S 2015*. New York, NY, USA: Association for Computing Machinery (pp. 317–320).

Moore, S., Nguyen, H., Bier, N., Domadia, T., & Stamper, J. (2023). Who writes tomorrow's learning activities? Exploring community college student participation in learnersourcing. In *Proceedings of the 17th international conference of the learning sciences*. International Society of the Learning Sciences.

Moore, S., Nguyen, H., & Stamper, J. (2022a). Participation and success with optional self-explanation for students in online undergraduate chemistry courses. In *Proceedings of the 16th international conference of the learning sciences*. International Society of the Learning Sciences (pp. 1381–1384). https://dev.stamper.org/publications/MooreShortICLS2022.pdf.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using gpt-3. In *Educating for a new future: Making sense of technology-enhanced learning adoption: 17th European conference on technology enhanced learning, EC-TEL* (pp. 243–257). Springer.

Moore, S., Nguyen, H. A., & Stamper, J. (2020). Evaluating crowdsourcing and topic modeling in generating knowledge components from explanations. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (pp. 398–410). Cham: Springer International Publishing.

Moore, S., Nguyen, H. A., & Stamper, J. (2021). Examining the effects of student participation and performance on the quality of learnersourcing multiple-choice questions. In *Proceedings of the eighth ACM conference on learning @ scale* (pp. 209–220).

Moore, S., Nguyen, H. A., & Stamper, J. (2022b). Leveraging students to generate skill tags that inform learning analytics. In *Proceedings of the 16th international conference of the learning sciences* (pp. 791–798).

Moore, S., Nguyen, H. A., & Stamper, J. (2023). Assessing the quality of multiple-choice questions using gpt-4 and rule-based approaches. In *Educating for a new future: Making sense of technology-enhanced learning adoption: 18th European conference on technology enhanced learning, EC-TEL*. Springer.

Moore, S., Stamper, J., Bier, N., & Blink, M. J. (2021). A human-centered approach to data driven iterative course improvement. In *Cross reality and data science in engineering: Proceedings of the 17th international conference on remote engineering and virtual instrumentation, vol. 17* (pp. 742–761). Springer.

Moore, S., Stamper, J., Brooks, C., Denny, P., & Khosravi, H. (2022). Learnersourcing: Student-generated content at scale. In *Proceedings of the ninth ACM conference on learning @ scale* (pp. 259–262).

Morales-Martinez, G., Latreille, P., & Denny, P. (2020). *Nationality and gender biases in multicultural online learning environments: The effects of anonymity*. New York, NY, USA: Association for Computing Machinery (pp. 1–14).

Nathan, M. J., Koedinger, K. R., Alibali, M. W., et al. (2001). Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the third international conference on cognitive science*.

Negi, S., Asooja, K., Mehrotra, S., & Buitelaar, P. (2016). A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the fifth joint conference on lexical and computational semantics* (pp. 170–178).

Ni, L., Bao, Q., Li, X., Qi, Q., Denny, P., Warren, J., Witbrock, M., & Liu, J. (2021). Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. https://doi.org/10.48550/arXiv.2111.10058.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*, 102–122. https://doi.org/10.1080/02602938.2013.795518.

OpenAI (2023). Gpt-4 technical report, arXiv:2303.08774.

Palisse, J., King, D. M., & MacLean, M. (2021). Comparative judgement and the hierarchy of students' choice criteria. *International Journal of Mathematical Education in Science and Technology*, 1–21. https://doi.org/10.1080/0020739X.2021.1962553.

Papanikolaou, K. A. (2014). Constructing interpretative views of learners' interaction behavior in an open learner model. *IEEE Transactions on Learning Technologies*, 201–214. https://doi.org/10.1109/TLT.2014.2363663.

Pirttinen, N., Kangas, V., Nikkarinen, I., Nygren, H., Leinonen, J., & Hellas, A. (2018). Crowdsourcing programming assignments with crowdsorcerer. In *Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 326–331). New York, NY, USA: Association for Computing Machinery.

Pirttinen, N., & Leinonen, J. (2022). Can students review their peers? Comparison of peer and instructor reviews. In *Proceedings of the 27th ACM conference on innovation and technology in computer science education, vol. 1*.

Pitt, R. (2015). Mainstreaming open textbooks: Educator perspectives on the impact of openstax college open textbooks. In *International review of research in open and distributed learning, vol. 16*.

Plak, S., van Klaveren, C., & Cornelisz, I. (2023). Raising student engagement using digital nudges tailored to students' motivation and perceived ability levels. *British Journal of Educational Technology*, *54*, 554–580. https://doi.org/10.1111/bjet.13261.

Polisda, Y. (2017). Peer review: A strategy to improve students' academic essay writings. *English Franca: Academic Journal of English Language and Education*, *1*, 45–60.

Quintana, R. M., Brooks, C., Smothers, C. V., Tan, Y., Yao, Z., & Kulkarni, C. (2018). Mentor academy: Engaging global learners in the creation of data science problems for MOOCs. Inernational Society of the Learning Sciences, Inc. [ISLS].

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. https://doi.org/10.48550/arXiv.2204.06125.

Rannikmäe, M., Holbrook, J., & Soobard, R. (2020). *Social constructivism—Jerome Bruner*. Cham: Springer International Publishing (pp. 259–275).

Reily, K., Finnerty, P. L., & Terveen, L. (2009). Two peers are better than one: Aggregating peer reviews for computing assignments is surprisingly accurate. In *Proceedings of the ACM 2009 international conference on supporting group work* (pp. 115–124).

Riggs, C. D., Kang, S., & Rennie, O. (2020). Positive impact of multiple-choice question authoring and regular quiz participation on student learning. *CBE—Life Sciences Education*, *19*, Article ar16. https://doi.org/10.1187/cbe.19-09-0189.

Rittle-Johnson, B., & Kmicikewycz, A. (2008). When generating answers benefits arithmetic skill: The importance of prior knowledge. *Journal of Experimental Child Psychology*, *101*, 75–81. https://doi.org/10.1016/j.jecp.2008.03.001.

Roberts, T. S. (2005). Computer-supported collaborative learning in higher education. In *Computer-supported collaborative learning in higher education* (pp. 1–18). IGI global.

Roitero, K., Barbera, D. L., Soprano, M., Demartini, G., Mizzaro, S., & Sakai, T. (2023). How many crowd workers do I need? On statistical power when crowdsourcing relevance judgments. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3597201.

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM conference on international computing education research V.1 (ICER 2022)*. ACM.

Scapin, D. L. (1982). Generation effect, structuring and computer commands. *Behaviour & Information Technology*, *1*, 401–410. https://doi.org/10.1080/01449298208914461.

Schmidt, F. A. (2013). The good, the bad and the ugly: Why crowdsourcing needs ethics. In *2013 international conference on cloud and green computing* (pp. 531–535). IEEE.

Singh, A., Brooks, C., & Doroudi, S. (2022). Learnersourcing in theory and practice: Synthesizing the literature and charting the future. In *Proceedings of the ninth ACM conference on learning @ scale* (pp. 234–245).

Singh, A., Brooks, C., Lin, Y., & Li, W. (2021). What's in it for the learners? Evidence from a randomized field experiment on learnersourcing questions in a MOOC. In *Proceedings of the eighth ACM conference on learning @ scale* (pp. 221–233). New York, NY, USA: Association for Computing Machinery.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology. Human Learning and Memory*, *4*, 592–604. https://doi.org/10.1037/0278-7393.4.6.592.

Snow, S., Wilde, A., Denny, P., & Schraefel, m. c. (2019). A discursive question: Supporting student-authored multiple choice questions through peer-learning software in non-stemm disciplines. *British Journal of Educational Technology*, *50*, 1815–1830. https://doi.org/10.1111/bjet.12686.

Snowball, J. D., & McKenna, S. (2017). Student-generated content: An approach to harnessing the power of diversity in higher education. *Teaching in Higher Education*, *22*, 604–618. https://doi.org/10.1080/13562517.2016.1273205.

Sommers, N. (1982). Responding to student writing. *College Composition and Communication*, *33*, 148–156.

Tackett, S., Raymond, M., Desai, R., Haist, S. A., Morales, A., Gaglani, S., & Clyman, S. G. (2018). Crowdsourcing for assessment items to support adaptive learning. *Medical Teacher*, *40*, 838–841. https://doi.org/10.1080/0142159X.2018.1490704.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*, 467–481. https://doi.org/10.1007/s10734-017-0220-3.

Tennant, J. P. (2018). The state of the art in peer review. *FEMS Microbiology Letters*, *365*, Article fny204.

VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*, 1–59.

Walsh, J. L., Harris, B. H., Denny, P., & Smith, P. (2018). Formative student-authored question bank: Perceptions, question quality and association with summative

performance. *Postgraduate Medical Journal*, *94*, 97–103. https://doi.org/10.1136/postgradmedj-2017-135018.

Wang, W., An, B., & Jiang, Y. (2020). Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems. *IEEE Transactions on Computational Social Systems*, *7*, 940–955.

Wang, X., Talluri, S. T., Rose, C., & Koedinger, K. (2019). Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *L@S 2019*. New York, NY, USA: Association for Computing Machinery.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, J., Hernandez-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., et al. (2021). Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 competition and demonstration track, PMLR* (pp. 191–205).

Wang, Z., Manning, K., Mallick, D. B., & Baraniuk, R. G. (2021). Towards blooms taxonomy classification without labels. In *International conference on artificial intelligence in education* (pp. 433–445). Springer.

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In *Artificial intelligence in education: 23rd international conference, proceedings, part I* (pp. 153–166). Springer.

Weir, S., Kim, J., Gajos, K. Z., & Miller, R. C. (2015). Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 405–416).

Wheeler, S., Yeomans, P., & Wheeler, D. (2008). The good, the bad and the wiki: Evaluating student-generated content for collaborative learning. *British Journal of Educational Technology*, *39*, 987–995. https://doi.org/10.1111/j.1467-8535.2007.00799.x.

Wiley, D., Bliss, T., & McEwen, M. (2014). Open educational resources: A review of the literature. In *Handbook of research on educational communications and technology* (pp. 781–789).

Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016). Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the third (2016) ACM conference on learning @ scale* (pp. 379–388).

Woolf, B. (2022). Introduction to IJAIED special issue, FATE in AIED, https://doi.org/10.1007/s40593-022-00299-x.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., et al. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology. General*, *143*, 804. https://doi.org/10.1037/a0033906.

Yeckehzaare, I., Barghi, T., & Resnick, P. (2020). Qmaps: Engaging students in voluntary question generation and linking. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14).

Zdravkova, K. (2020). Ethical issues of crowdsourcing in education. *Journal of Responsible Technology*, *2*, Article 100004. https://doi.org/10.1016/j.jrt.2020.100004.

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, *10*, 541–552. https://doi.org/10.14778/3055540.3055547.