Toward Automation to Support Creation and Evaluation of Pedagogically Valid Multiple-Choice Question Assessments at Scale

Steven Moore

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Human-Computer Interaction at Carnegie Mellon University

CMU-HCII-24-109
September 2024

Thesis Committee:
John Stamper (Chair), Carnegie Mellon University
Ken Koedinger, Carnegie Mellon University
Sherry Tongshuang Wu, Carnegie Mellon University
Christopher Brooks, University of Michigan

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University

Acknowledgments

You know who you are <3

Abstract

Multiple-choice questions (MCQs) are the predominant form of assessment in educational environments, known for their efficiency and scalability. Traditionally, these questions are crafted by instructors, a method that despite its expertise often results in inconsistencies and errors. In response to these limitations and the need for scalability, learnersourcing has been leveraged, which involves students in the question creation process. Although this method capitalizes on the diverse perspectives of students, it also leads to significant variability in the quality of the questions produced. Additionally, while recent advances in artificial intelligence have facilitated more scalable and automated methods for generating MCQs, these AI-driven methods still suffer from many of the same shortcomings as those created by humans. Current evaluation methods for MCQs predominantly rely on human judgment, which introduces subjectivity and lacks scalability. While automated evaluation methods provide scalability, they fall short in adequately assessing the educational value of questions, focusing instead on surface-level features that do not match expert evaluation.

In this thesis, I explore various methods for creating and evaluating educational content, grounded in learning science research and guided by the use of rubrics. I demonstrate that students, with minimal scaffolding and technological support, are capable of generating high-quality assessments. I have also investigated the potential of involving both students and crowdworkers in the generation and evaluation of the skills required to solve problems. Building on this, we developed a new method that leverages LLMs to enhance the efficiency and accuracy of these processes. Furthermore, I have shown that crowdworkers can effectively use rubrics to evaluate questions with a level of accuracy comparable to human experts. Through these crowdsourcing and learnersourcing studies, I examine how specialized knowledge and expertise influence the success of content creation and evaluation. This work culminates in the proposal of the Scalable Automatic Question Usability Evaluation Toolkit (SAQUET), a new standardized method for evaluating educational MCQs.

This work contributes to the fields of educational technology, learning sciences, and human-computer interaction. By harnessing the capabilities of crowdsourcing, learnersourcing, and generative AI, this research demonstrates how the generation and evaluation of educational content can be vastly improved. It introduces a standardized approach to assessment processes, enhancing the quality and consistency of educational evaluations across various domains. By providing a scalable framework that leverages advancements in generative AI, this work propels the field of educational technology forward, addressing critical challenges related to the creation and evaluation of assessments. Ultimately, these contributions offer a foundation for future innovations in educational content development and quality assurance.

Contents

Acronyms

MCQ	Multiple-Choice Question
SAQ	Short Answer Question
IWF	Item-Writing Flaws
NLP	Natural Language Processing
LLM	Large Language Model
ML	Machine Learning
AQG	Automatic Question Generation
KC	Knowledge Component
KCM	Knowledge Component Model
CTA	Cognitive Task Analysis
IRT	Item Response Theory
RMSE	Root Mean Square Error
AMT	Amazon's Mechanical Turk
SAQUET	Scalable Automatic Question Usability Evaluation Toolkit

List of Figures

List of Tables

Chapter 1: Introduction

Assessments are foundational to education, but creating high-quality assessments is a challenging endeavor. Multiple-choice questions (MCQs) are the most commonly used form of assessment, relied upon extensively due to their efficiency and scalability [63]. Despite their widespread use, even experts in the field continue to struggle with the process of designing effective MCQs. Experts may face difficulties in ensuring that questions accurately measure comprehension without ambiguity, crafting distractors that are plausible yet clearly incorrect, and aligning each question with specific learning objectives, all of which require an understanding of both the subject matter and effective pedagogical strategies [105].

Despite the expertise typically required for generating MCQs, students in various online courses have been employed to create these assessments, thereby mitigating potential expert blind spots. This is known as learnersourcing, where students engage in an educational activity that produces data which can be used by future learners [110]. Numerous systems have been developed to facilitate this process, enabling students to generate millions of questions, including MCQs and short-answer questions [58, 109, 192, 227, 237]. However, while learnersourcing taps into a diverse human resource, it introduces its own challenges related to maintaining question quality and managing the time required for question generation.

Over the past decade, automatic methods for creating educational MCQs using natural language processing (NLP) techniques have significantly improved, enabling a mass production of MCQs across any domain [170]. The advent of Large Language Models (LLMs) has particularly enhanced the ease of generating assessments, especially MCQs. These automated methods vary in approach; some require a document to provide context, while others operate with just a simple prompt, enabling a pretrained model to generate questions [75]. This flexibility has made the process of creating MCQs more accessible and efficient than ever before.

While various methods of MCQ creation exist, including those by experts, students, and AI, each approach has its own unique set of flaws as well as some that overlap. For example, an expert instructor may unintentionally provide extra details in the correct answer, thereby signaling its correctness. Students might create nonsensical distractors, and AI-generated questions could include unnecessary information in the question stem. Furthermore, questions generated by any method often focus on lower levels of Bloom's Revised Taxonomy, predominantly testing recall rather than higher cognitive processes which would be more desirable [93, 112]. Additionally, regardless of the creation method, the alignment of questions with specific skills or knowledge components frequently goes unaddressed. Efforts to improve this through learnersourcing the knowledge components have been made; however, student attempts often yield poor outcomes. While crowd-sourced knowledge component tagging offers some improvement, it still falls short of ideal, highlighting the pervasive limitations across all current MCQ creation methods [161].

Current methods for evaluating the quality and the skill or knowledge component tags of assessments rely heavily on expert human judgment, which is both subjective and challenging [122]. Traditional tools like psychometrics and item response theory (IRT) also come with drawbacks; they require extensive student data and risk exposing learners to poorly designed or potentially harmful questions [18]. While crowd-sourcing evaluations provide a more robust alternative, this method is time consuming and expensive. Automated approaches, although widely used, offer only superficial analysis and lack depth, often ignoring the pedagogical implications of the questions. Consequently, despite their widespread use, we lack effective measures of assessment quality that are standardized and relevant across different domains. This gap highlights the need for more reliable and efficient evaluation methods in educational assessment.

Addressing these evolving needs, I developed a new technique in educational assessment that automatically applies the Item-Writing Flaws (IWF) rubric, which contains a set of criteria used to evaluate the quality of educational MCQs [196, 221, 222]. This domain-agnostic method systematically evaluates MCQs using the verified IWF rubric, and I demonstrate its effectiveness across a variety of subject areas in higher education. The method, along with a related LLM-based application of a SAQ rubric, is further evaluated against human assessments across varying levels of expertise. Additionally, I developed a complementary method for generating and associating skill tags with assessment items. This dissertation details the development and effectiveness of these methods, showcasing their application through empirical studies and their integration into existing educational frameworks. By automating the evaluation process and providing a robust set of criteria, these techniques address the limitations of subjective human judgment and the lack of standardization in traditional assessment methods. Designed for ease of use and broad applicability, these methods offer a transformative approach for educators and institutions aiming to enhance the reliability and validity of their assessments.

Ultimately, the broader implications of this work are significant, making advances in learnersourcing, educational crowdsourcing, skill tagging, and the creation and evaluation of educational content. This work not only contributes to the academic community but also has practical implications, enhancing how educators and institutions assess learning outcomes using evaluated assessment items.

1.1 Thesis Statement

With the proliferation of methods for creating MCQs, including experts, novices, students, and AI, generating these questions has become more accessible than ever before. However, these methods are far from perfect, highlighting the need for a gold standard in evaluating question quality. This standard must surpass the limitations of superficial automated methods and the inefficiencies of time-consuming subjective evaluations, with a strong emphasis on the pedagogical impact of the questions. Furthermore, there is a need to organically improve question construction by providing actionable feedback through this evaluation process. High-quality questions should be accessible to everyone, valid, and reliable. These questions should also include knowledge component data to support adaptive learning systems and learning analytics.

The core hypothesis of this research posits that it is possible to empower individuals, regardless of their expertise or background, to create and refine high-quality educational assessments. This work demonstrates how MCQs can be automatically evaluated against pedagogically aligned criteria accurately and effectively. By implementing a domain-agnostic approach, our method not only improves the quality of MCQs, but also generates the low-level skills the question assesses. This evaluation process not only enhances the assessments themselves, but also elevates the capabilities of the authors, fostering better question creation across diverse educational contexts.

1.2 Thesis Overview

This proposal progresses through chapters detailing related work (Ch. 2), learnersourcing assessments (Ch. 3, Ch. 4), knowledge component tagging (Ch. 5, Ch. 6), equitable participation in these tasks (Ch. 7), question evaluation using crowdsourcing (Ch. 8), automatic MCQ evaluation (Ch. 9, Ch. 10) comparison of crowdsourcing, learnersourcing and AI rubric applications (Ch. 11), generation and associate of skills (Ch. 12) discussion and future work (Ch. 13) and a conclusion (Ch. 14), organized as follows:

In Chapter 2, I review literature on crowdsourcing in educational settings to illustrate how collective intelligence shapes learning material creation and evaluation. I then focus on learnersourcing, a specific type of crowdsourcing that engages students in question generation. This leads to an analysis of what defines an effective educational question, highlighting the involved knowledge components and cognitive processes. I introduce the Item-Writing Flaws rubric, commonly used to analyze MCQs, and compare different methods for evaluating educational content, contrasting human judgment with automated systems. The chapter concludes by exploring emerging trends in human-AI collaboration within educational contexts.

In Chapters 3 and 4 I explore how students create MCQs and short-answer questions, analyzing the range of qualities and characteristics that may influence their creation. My findings reveal that students are capable of producing high-quality questions even with minimal technological support and scaffolding; however, many questions still exhibit significant flaws. This demonstrates the potential for students to contribute effectively to question creation, even when participation is optional and without extensive systemic support. I examined the need for enhanced evaluation processes to identify high-quality questions from student contributions and guide not only students, but also other non-experts, in developing superior educational assessments. These chapters highlight the role of effective evaluation techniques in recognizing and fostering quality in educational content creation, setting the stage for subsequent discussions on proposing new methods for evaluating and refining educational assessments.

In Chapters 5 and 6, I explore two approaches aimed at improving the knowledge component tagging process for educational assessments, utilizing crowdsourcing and learnersourcing methods. Initially, the crowdsourcing approach proved ineffective, despite being scaffolded and supported with examples, prompting a reevaluation and reformulation of the task into a compare and contrast activity. Subsequently, this revised task was tested using students in the context of several online courses with more expertise to assess the impact of increased knowledge on the process. However, the results remained suboptimal, indicating that this method may not be an efficient use of students' time or knowledge given the specific task. These findings indicate the necessity of exploring alternative approaches that could better harness domain expertise or technology for more effective knowledge component tagging.

In Chapter 7, I investigate the demographics and performance levels of students participating in learnersourcing activities to assess inclusivity. I conducted a study involving a MCQ generation activity across three courses at two community colleges, analyzing how students' demographic data and performance influenced their participation in these optional tasks. The results indicated that students with higher scores in formative and summative assessments were more likely to engage in learnersourcing. However, most of the top 10% scorers did not participate, likely viewing the activities as unnecessary. This pattern suggests that while many students are willing to engage, the highest achievers who could provide the most insightful contributions often abstain. These findings highlight the need for strategies to foster wider and more diverse participation in learnersourcing activities to maximize their educational impact.

In Chapter 8, I explore a crowdsourcing approach to evaluating the quality of MCQs in the domains of mathematics and chemistry. This study involved training crowdworkers to use the IWF rubric, assessing their ability to apply it to questions outside their areas of expertise. The findings indicate that the majority of crowdworkers, despite limited domain knowledge, could accurately apply the IWF rubric, suggesting that the rubric's scaffolding and guidelines are sufficiently robust. While this method shows promise for an automated and scalable evaluation system, it also presents challenges, including significant time and cost requirements. Nevertheless, the effectiveness of the IWF rubric in aiding accurate question evaluation across multiple dimensions opens the possibility for its implementation through programming techniques or LLMs, enhancing its utility and scalability.

In Chapters 9 and 10, I detail the development and evaluation of the Scalable Automatic Question Usability Evaluation Toolkit (SAQUET), which utilizes a combined rule- and LLM-based approach to automatically apply the 19 criteria of the IWF rubric, as discussed in Chapters 3 and 8. This toolkit incorporates advances in NLP to analyze components of MCQs and identify potential flaws. It offers a domain-agnostic and scalable solution for evaluating educational MCQs against these established criteria, overcoming shortcomings I identified in other commonly used evaluation metrics. Through testing on questions from a variety of distinct academic domains, we demonstrate SAQUET's accuracy and effectiveness. The results indicate that this approach could significantly enhance the accuracy and comprehensiveness of quality evaluations for educational assessments, addressing some of the evaluation and question generation challenges previously identified.

In Chapter 11, we compare the effectiveness of applying the IWF rubric and a 9-item SAQ rubric to questions from three distinct domains. The comparison involves two forms of crowdsourcing, varied by expertise level, three popular LLMs, SAQUET, and expert evaluation. We assess how these different methods successfully apply the two rubrics for evaluating question quality. Our findings indicate that the automated methods achieved near-perfect accuracy for certain criteria, while human evaluations excelled in accuracy for others. These results suggest that a hybrid approach, using automated methods to evaluate the more reliable criteria and relying on human evaluators for the remaining ones, could be a promising and scalable solution for improving question quality evaluation.

In Chapter 12, I introduce a new method for unsupervised skill generation and association with educational MCQs, eliminating the need for additional data such as contextual information or instructional text. In the domains of Chemistry and E-Learning, I demonstrate that KC labels generated using an LLM-based approach matched expert-generated ones approximately 50% of the time. However, when evaluated by three domain experts in each subject area, the LLM-generated KCs were significantly preferred over the expert-generated ones. This suggests that the method may outperform traditional human-generated mappings or, at the very least, provides a strong foundation for initial skill mapping that can be further refined by experts.

In Chapters 13 and 14, I discuss the nine key contributions of this thesis. I detail the implications of this work and the future directions we can tackle from here. Lastly, I present a conclusion that summarizes all of this work.

1.3 Summary of Contributions

This dissertation makes four significant contributions to the fields of learnersourcing, crowdsourcing, educational assessment, and technology-enhanced learning. First, it demonstrates through three learnersourcing studies how students from various domains and higher education institutions can generate high-quality questions with minimal technological support, not requiring a specialized tool. It also provides an analysis of student participation in these activities from an equity perspective, emphasizing the necessity for more scalable and effective evaluation techniques. Second, this work shows how varying levels of domain expertise and experience can impact the evaluation of educational content. While rubrics can scaffold some of these processes, advanced domain knowledge may be required in some form. Third, the research introduces innovative approaches for developing knowledge component tags for assessment items by leveraging crowdsourcing, learnersourcing, and LLMs. These methods explore a scalable solution to enhancing assessment items by generating and associating essential metadata. Fourth, this work presents the first automated and scalable toolkit for the evaluation of MCQs that accounts for the pedagogical implications of questions and operates across various domains without necessitating advanced expertise. This development promises to significantly enhance the quality and efficacy of question evaluation processes in educational settings.

Chapter 2: Background

This section commences with a literature review on crowdsourcing within educational contexts, elucidating how collective intelligence can be harnessed to develop and assess learning materials. Subsequently, I explore the nuanced concept of learnersourcing, a derivative of crowdsourcing that specifically leverages student engagement, for question generation purposes. The discourse then shifts to identifying the attributes of an effective educational question, examining both the knowledge components it measures and the cognitive processes it stimulates. The following part introduces the Item-Writing Flaws rubric, a prevalent tool for critiquing multiple-choice questions. Thereafter, the discussion extends to an examination of methodologies for educational content evaluation, contrasting human judgment with automated approaches. The section culminates by examining the nascent integration of human-AI collaboration in educational settings.

2.1 Crowdsourcing in Education

In educational settings, crowdsourcing has emerged as an innovative approach to enhance question generation, refine content through feedback, furnish hints, and categorize the foundational knowledge required for answering questions accurately [1, 77, 106, 163]. However, a significant challenge persists: improving the quality of contributions without sacrificing the broad-scale applicability inherent to crowdsourcing [232]. Given that such tasks frequently demand specific expertise, ranging from subject matter proficiency to pedagogical understanding, crowdworkers may find themselves at a disadvantage [113]. Addressing this issue, research has shown that equipping crowdworkers with expert-crafted examples and detailed rubrics not only bolsters the quality of their work but also diminishes the time invested and mitigates the complexities associated with insufficient expertise [64, 65]. An illustrative study revealed that when crowdworkers without pedagogical training were given a rubric to evaluate writing, their assessments aligned closely with those of seasoned educators [5]. Employing rubrics thus fulfills a dual function in crowdsourcing initiatives: it prescribes uniform standards for evaluation and adeptly taps into the rich tapestry of perspectives offered by a diverse crowd.

2.2 Learnersourcing

Learnersourcing involves students engaging in activities that produce content which can be leveraged by future learners [110]. It has been used in many online courses across a variety of domains, where students are typically tasked with generating questions, making hints, or providing feedback [107]. Having students generate short answer questions or MCQs that can then be used as practice opportunities in the current or future courses is a particular focus of much learnersourcing research [227, 237]. An obvious challenge that arises from optional activities is getting the students to participate with them and making a meaningful contribution [37]. Previous research has demonstrated that completing optional course activities is strongly related to a student's performance in a course [117]. As researchers and educators, we want students to participate in learnersourcing activities, given that such activities can provide useful learning data, contribute to the instructor's assessment question banks, and benefit student learning [4]. However, it is important to understand what factors might influence students's decision to participate in these optional activities, as addressed in the study by [213]. To determine if such activities are reaching all students in the course, or only those from the commonly represented demographics or top-performing group, we need to investigate these factors as they relate to students contributing to these learnersourcing tasks.

2.2.1 Student Engagement in Online Courses

Online courses offer students different affordances compared to traditional in-person ones, which can be both beneficial and detrimental to learning depending on the student. A study by [197] found that over 90% of the students enrolled in online computer science courses participated at least once, but overall participation rates ranged along a continuum from active to passive participation. They found that student participation within these courses varied by demographics, such as ethnicity and age. Particularly in STEM, evidence suggests that online courses can perpetuate enrollment and participation gaps for women or ethnicities that are traditionally underrepresented in these courses [114].

Student engagement with an online course can be defined by their participation in its learning activities [82]. Multiple studies have linked student performance to their engagement with the course materials, indicating that students who actively participate and do more activities have a higher chance to pass the course and receive a higher grade [34]. While research supports the benefits of having students participate in optional activities found in online courses, other factors such as the demographics of the students may also influence their participation and ultimately their success in the course [193]. For instance, student motivation in STEM courses can be affected by stereotype threat, causing a lack of a sense of belonging [23]. This lack of participation, particularly when it involves learnersourcing, presents several challenges that propagate throughout the course. When students have lower levels of participation, they do fewer activities, which can pose difficulties in modeling their learning [138]. Students doing fewer activities also leads to less data being generated, which can hinder the efficacy of instructional interventions, such as recommending practice problems [13].

2.2.2 Question Generation

Previous work has utilized learnersourcing techniques to have students generate short answer questions and MCQs [43, 237]. For the creation of short answer questions, [43] found that the process is beneficial to student learning, as it increases their engagement with the material and invokes critical thinking. For the creation of MCQs, [237] found that a majority (86%) of the student- generated questions met their quality threshold and identified several social features, such as question ownership, that kept students motivated to make contributions. It is not typical that learnersourced contributions achieve such a high quality, even when the students are trained prior to making a contribution to the task [150]. In addition to providing training, learnersourcing activities are commonly presented via a separate tool or embedded within high-stakes assessments, to improve the quality and increase the participation on the task [77, 99]. One such popular system is PeerWise, which provides students with a custom learning environment for collaboratively generating and sharing questions [58]. Previous research has demonstrated that students authoring questions in the PeerWise system has had positive effects on student learning and improved their performance on exams [70, 149].

The success of such systems is likely a result of how the students' generation of MCQs has been proven to positively impact their deep learning [57, 61, 70]. Another system, RiPPLE, also enables students to generate MCQs and formulate distractors for them, which requires them to think deeply about potential misconceptions [109]. The study involving RiPPLE found that students using the system felt positive about their experience, which ultimately led to measurable learning gains. An average of 1.6 questions per student were authored during their use of the system over a five week period in their course. Other learnersourcing approaches, such as Upgrade, take an offline approach, that generates learning opportunities from prior student solutions to open-ended problems [227]. They found that students achieved the same learning outcomes in 30% less time using Upgrade-created questions instead of traditional open-ended ones.

2.3 KCs and Cognitive Processes

When creating multiple-choice questions, it's important to incorporate several key elements to enhance their educational value. In addition to providing feedback and hints, which guide students towards understanding the correct answers, the questions should be designed to clearly align with specific knowledge components or competencies it is intended to assess. This alignment helps in accurately measuring student proficiency in targeted areas and aids educators in identifying gaps in knowledge or understanding. Additionally, the questions should cover a range of Bloom's Revised Taxonomy levels. This ensures they assess various cognitive skills from basic recall of facts to higher-order thinking like analysis and evaluation.

2.3.1 Knowledge Components

A knowledge component is a specific piece of knowledge required to address a particular problem within a digital learning environment [116]. These knowledge components enable various learning analytics tools, such as learning curves, open learner models, and adaptive learning systems [216]. Studies have shown that digital learning environments utilizing adaptive systems powered by these knowledge components, like cognitive tutors, allow students to achieve mastery 26% faster than traditional classroom instruction [118]. The effectiveness of these systems depends on the precise definition and application of knowledge component tags that accurately represent the knowledge needed to solve specific problems.

In learning analytics, knowledge components are typically more detailed than broader learning objectives or standards, which are typically linked to course sections encompassing several skills [19]. For example, a learning objective in an algebra course might be "Graph linear and quadratic functions and show intercepts, maxima, and minima," which covers multiple concepts. In contrast, a knowledge component might be narrowly defined as "Identify the slope from an equation in the form of y=mx+b," focusing on a single problem within those broader objectives. Knowledge components must be relevant to both the problem and the course context, avoiding assessment of overly basic knowledge such as the meaning of "+" and "=" in an algebra course.

Knowledge component tagging is traditionally performed by experts who understand the necessary granularity and the subject matter [115]. Often, this involves a cognitive task analysis (CTA) or think-aloud process, where an expert articulates their thought processes during task performance, helping to identify and record the knowledge components required [44]. While this method produces an accurate knowledge component mapping and effective instructional designs, it requires significant effort [128]. Moreover, it helps domain experts avoid the "expert blind spot," a phenomenon where experts might overlook basic steps that have become intuitive, potentially complicating the knowledge component tagging process [171].

2.3.2 Automated KC Generation and Evaluation

The growing limitations of manual KCM construction, which relies exclusively on human input, underscore the need for more effective approaches, as manual methods often demand significant resources and time. Automated approaches for generating KCMs have emerged as valuable tools that can enhance, rather than replace, human efforts [20]. These methods employ data-driven approaches with varied human intervention, like Learning Factors Analysis (LFA) and Q-matrix inspection, to categorize questions under existing KCs within a predefined search space in educational software [21, 38, 211]. Specifically, LFA relies on human input to suggest factors that could explain task difficulty or the transfer of learning between tasks. The results from these models require assigning human readable labels after the fact, which has been found to be a pain point [134].

A KCM can be leveraged to predict student performance, the accuracy of which can reflect how well the KCM represents the knowledge students gain from different educational activities [139]. One such method, the Additive Factors Model (AFM), is a logistic model that tracks students' knowledge growth by observing changes in their performance across repeated practice with targeted KCs [72]. The AFM builds on IRT models by incorporating an underlying KC model and considering the learning that happens as students repeatedly apply each KC [205].

Automated approaches for generating KCMs that operate largely without human intervention typically follow two main strategies: generation or classification [173]. In terms of generation, significant efforts focus on creating knowledge graphs or extracting concepts from digital textbooks, in addition to deriving KCs from student performance data [40], for example via matrix factorization [24, 62] and VAE-based methods [176]. However, these methods often face challenges related to interpretability, not only due to the opaque nature of the algorithms used, but also because the generated labels may not hold meaningful insights for educators [210]. On the classification front, the goal is to assign existing KCs to problems based on semantic information contained in the assessment text, a process that has proven effective in domains like Math and Science [182, 224]. However, for areas without a well-defined standard or an established bank of KCs, such as those outside the common core standards, this classification approach presents a significant challenge due to the absence of predefined labels for categorization [130]. Another related problem is establishing the equivalence of individual KCs across different learning platforms which often use varying nomenclature to refer to the same learning objectives. Prior work explored the application of machine translation techniques that consider assessment context and textual content to identify equivalent KC pairings [131].

Notably, models developed or improved through automated or semi-automated techniques frequently surpass their manually constructed counterparts, especially in predicting student performance [7]. For example, evidence from previous research shows that a KCM refined through a combination of human judgment and automated methods can enable students to achieve mastery 26% faster [118].

2.3.3 Bloom’s Revised Taxonomy

It is beneficial for student learning if they encounter a variety of MCQs that target higher-order cognitive processes according to Bloom's Revised Taxonomy [93]. This taxonomy consists of six hierarchical categories, each representing the cognitive processes required to answer the question, ranging from recalling information to creating new patterns or structures [119]. Previous research has shown that MCQs commonly assess lower-level cognitive processes, such as recall, but they can assess all levels [55]. Assigning a Bloom's Revised Taxonomy label to each question can improve problem selection and learning analytics [49]. Automated methods for determining the cognitive level of questions have shown promise, with accuracy as high as 84% compared to human labels [158]. However, these methods often require large amounts of training data or expert time, making them inaccessible and difficult to scale.

2.4 Item-Writing Flaws

Developing MCQs that cover the appropriate concepts and target higher cognitive levels can be challenging, even for expert instructors [68]. To assess the quality of MCQs, different item response theory and statistical methods have commonly been utilized [63, 105]. These methods often use collected student data, which details if their choice was correct, which distractor(s) they selected, and how many attempts they took to answer the question correctly. However, testing and assessing MCQs in this manner poses a potential problem if the questions are poorly constructed, as they can negatively impact students’ performance and achievement [46]. To help prevent these negative effects, previous studies have relied on qualitatively reviewing MCQs prior to testing them with students to confirm their validity [12, 31, 222]. These studies often evaluate the questions using a series of guidelines, such as the popular item-writing flaws (IWF) guideline that provides a validated rubric consisting of 31 unique items for assessing the quality of an MCQ [86].

Many studies have made use of the IWF guidelines, either by adopting the original 31-item rubric or creating an abridged version for their own purpose, as some of the items are not always applicable to the questions in a particular domain [31, 196, 222]. These studies often include an evaluation of the cognitive levels the MCQ assesses, which traditionally are recall and comprehension [207, 243]. One particular study assessed the quality of over two thousand instructor generated MCQs by utilizing a 19-criteria version of the IWF guidelines [221]. They had several reviewers analyze the MCQs for IWFs and evaluate the cognitive level the question assesses as either recall or application. Ultimately they found that nearly half of the questions were deemed unacceptable due to containing too many IWFs. The present study utilizes the same 19-IWF guidelines and criteria for assessing MCQs at the recall or application cognitive level from [221].

However, while previous work focuses on applying the guidelines to instructor-generated questions, we apply them to student-generated ones.

2.5 Educational Content Evaluation

Previous research has identified flaws in MCQs across various domains and teaching levels, including high-stakes standardized tests developed by psychometricians and domain experts [221]. These MCQs often find repeated use in test banks, practice sets, and training materials over the years. Consequently, there is a need for ongoing quality evaluation of these pre-existing questions, not just newly generated ones. This can complement analyses based on student performance data, such as those offered by Item Response Theory (IRT) [18]. However, evaluating MCQs before their implementation is crucial to avoid exposing poorly designed questions to learners, which can impede their learning [184]. Crafting high-quality questions remains a significant challenge, and evaluating their quality poses an even greater one, demanding consistency, scalability, and consideration of the questions' application contexts.

2.5.1 Automated Evaluation Methods

Over the past decade, automated MCQ quality evaluation has relied on metrics such as BLEU, METEOR, and ROUGE [170]. These metrics primarily assess similarity to a gold standard without considering educational value or effectiveness in evaluating student knowledge [153]. While previous research states these “standardized” metrics facilitate comparison across studies, they involve numerous hyper-parameters that can vary by task and are often insufficiently reported, complicating precise comparisons and replications [144]. Moreover, prior work has demonstrated that these metrics do not sufficiently align with human evaluation [144, 228]. To align more closely with human evaluation while maintaining scalability, alternative automated approaches have explored metrics like perplexity, diversity, grammatical error, complexity, and answerability [189, 230]. These have been applied to both machine- and human generated questions, offering a broader evaluation that extends beyond mere readability to include aspects critical for educational assessments.

When evaluating MCQs, perplexity assesses a language model's ability to predict question and answer text based on its training data [35]. Lower scores suggest more coherent questions and answers with predictable language patterns, whereas higher scores indicate complexity or atypical text, suggesting the questions could be unclear or poorly structured. Diversity evaluates the range in vocabulary, structure, and content across generated texts, ensuring a variety of questions and answers and reducing repetition [129]. A higher diversity score indicates greater uniqueness among MCQs, avoiding repetitive phrases and templated patterns. Grammatical errors pinpoint grammar violations, such as incorrect verb tense or spelling, quantified for each MCQ.

Complexity is typically assessed through cognitive complexity, using Bloom's Revised Taxonomy to assign difficulty levels to MCQs based on the cognitive skills required to answer them [122]. Bloom's Revised Taxonomy categorizes cognitive skills ranging from recall (remembering) to higher-order skills (creating), with questions demanding higher-order thinking deemed more cognitively complex [74]. Answerability measures how accurately a question can be answered, using the provided context or common knowledge. Recently, LLMs such as GPT-4 have been used to automate this evaluation metric [189]. Specifically, the Prompting-based Metric on ANswerability (PMAN) strategy employs three prompts to evaluate a question's quality by how well an LLM can answer it, demonstrating that it aligns with human judgments [228].

2.5.2 Human Evaluation Methods

Despite the advancement of automated methods for evaluating multiple-choice question (MCQ) quality, human evaluation remains the gold standard due to its accuracy, although it is often subjective and based on intuitive metrics like "difficulty" or "acceptability" [122, 125, 170]. These human assessments, involving expert judgment of the questions against "best practice" conventions, are challenging to standardize, replicate, and scale due to their time-intensive nature [86]. Typically, experts or instructional designers use a standardized rubric to assess both automatically generated and student-generated MCQs, helping to decrease subjectivity and enhance the reproducibility of the evaluations [32, 122, 186].

To further assess the quality of student-generated questions, researchers have used IRT techniques analyzing student performance data, which, if not pre-evaluated for quality, might detrimentally affect student outcomes due to potentially poor question construction [46, 122]. Alternatively, experts and peers frequently review these questions using a detailed rubric that includes criteria like language coherence, correctness, and perceived difficulty. However, past evaluations have often overlooked deeper pedagogical aspects, such as how well questions integrate into the course or assess previously unexamined content [22, 125].

2.6 Human-AI Partnerships

Partnerships for co-creating educational content typically follow four key phases: creation, evaluation, utilization, and instructor/expert oversight [108]. By integrating advances in large language models (LLMs), learnersourcing can be enhanced with AI, providing students with nearly instant feedback on their creations and improving the quality of their contributions [60]. These collaborations between students, instructors, and AI offer extensive opportunities for both creating and evaluating content [212]. Advances in natural language processing (NLP) and generative models enable AI to play a critical role in co-creating content and in automating its quality evaluation. Learning analytics can also support the evaluation by analyzing student performance on AI-co-created assessments versus traditional ones. Existing research has investigated the effectiveness and innovation of AI-generated learning resources [202], utilizing NLP [161], trust-based networks [54], and deep learning [174] to help evaluate both student- and AI-generated content. Although human input is essential, there is a growing need to further leverage AI to support students and instructors in developing educational content.

Chapter 3
Students Generating High Quality MCQs

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Huy Anh Nguyen, and John Stamper. "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. 2021.

3.1 Introduction

Multiple choice questions (MCQs) are a popular form of both formative and summative assessment, widely used in higher education, and often accounting for a considerable portion of a student's course grade [63, 147]. MCQs are advantageous because they are efficient to score, can be graded objectively, enable item-analysis calculation upon student completion, and require less time for students to respond [36, 83]. While MCQs traditionally assess students for recall and comprehension, they can also probe for higher-level cognitive processes such as the knowledge application and problem analysis [112, 149, 207]. In addition to evaluating student knowledge in both low-stakes and high-stakes environments, MCQs offer a scalable and equitable means of assessment [192]. The need for such scalability in assessment continues to increase, as class sizes continue to grow and more educational materials shift to being online [68]. With traditional authoring techniques for creating MCQs, teachers will be challenged to keep up with increased demand for new and quality assessments, making a more scalable solution desirable.

Instructors and teaching staff rarely have the time or incentive to develop quality MCQs for formative assessment; instead their efforts are often focused on creating high-stakes assessments such as quiz or exam questions [100, 172]. The continual creation and improvement of MCQs allows for a greater breadth of topic coverage, helps to identify well constructed and valid assessments, and as a result, enables improved learning analytics. However, creating MCQs presents an issue of scalability, which recent efforts have tried to improve by enlisting students in the process of MCQ generation, known as a form of learnersourcing, to varying degrees of success [87, 237]. Learnersourcing is a form of crowdsourcing in which students contribute novel content for future learners while engaging in a meaningful learning experience themselves [227]. While platforms like PeerWise [58], Quizzical [192], and RIPPLE [109] utilize learnersourcing by allowing students to author MCQs, they are not directly integrated with the instructional content and accompanying activities, requiring students to change between tools and invest ample time into the process of authoring even a single question. Students and their data are being leveraged to create assessments, but we need to better utilize them in this process, amplifying their voice and viewpoints, without detracting from their learning or requiring an excessive amount of their time. Previous work indicates that the process of having students generate MCQs can benefit their learning [4]. By better understanding how students participate and interact with generating MCQs, we can work towards improving the process so that it benefits both the student's learning experience and the quality of the questions they create.

In order to discern how students engage in the MCQ generation process, we sought a solution that does not require an additional tool or interrupt the context of their instruction. In particular, we deployed a completely optional MCQ generation activity in the context of seven instances of an online course. Students working through the course, consisting of multiple pages of instructional content and assessments, were presented with low-stakes activities that were optional to complete as they worked throughout the course. We investigated how this elicitation of having students generate an MCQ, given that it was optional, presented directly among the course context, and surrounded by the accompanying instructional text, would garner participation for the activity. From the student contributions collected, we evaluated the quality of the MCQs, determining if they were acceptable or contained certain item-writing flaws. The student-generated MCQs were also assessed for their cognitive level, in particular based on whether they targeted the typical recall level or if they extended to the higher level of application and analysis [207, 240]. Finally, we explored how different aspects of student interaction in the online course, such as their performance on other low-stakes activities, correlate to the quality of the MCQs they generated.

Through the investigation of these research questions, our work makes the following contributions towards learnersourcing. First, our experimental results suggest a set of student behaviors that influence their participation in an optional learnersourcing task. Secondly, the study demonstrates that students can provide recall- and application-level multiple choice questions, without training or scaffolding. Third, we identified features of student performance in an online course that are correlated to the quality of the multiple-choice questions they generate.

3.2 Methods

3.2.1 Study Context and Students

For this study, we used data collected from seven instances of the same introductory chemistry course being taught at a community college on the west coast of the United States. This course provides students with fundamental knowledge of chemistry concepts, preparing them for future biology and chemistry courses. There are no prerequisites for the course, outside of having prior experience with intermediate algebra, which most of the students had from high school. Additionally, the course is generally geared towards freshman and sophomore undergraduate students from varying degree backgrounds, with a majority of the students pursuing a chemistry-related degree, such as a bachelor's in biochemical engineering. The collected data we used comes from the summer and fall semesters of 2020, when the introductory chemistry course was offered in the OLI system. A single instance of the course was taught during the summer semester and the remaining six instances were taught during the fall 2020 semester. A further breakdown of the course offerings, including the anonymized instructor, semester, and number of students that accessed the course materials can be found in Table 3.1.

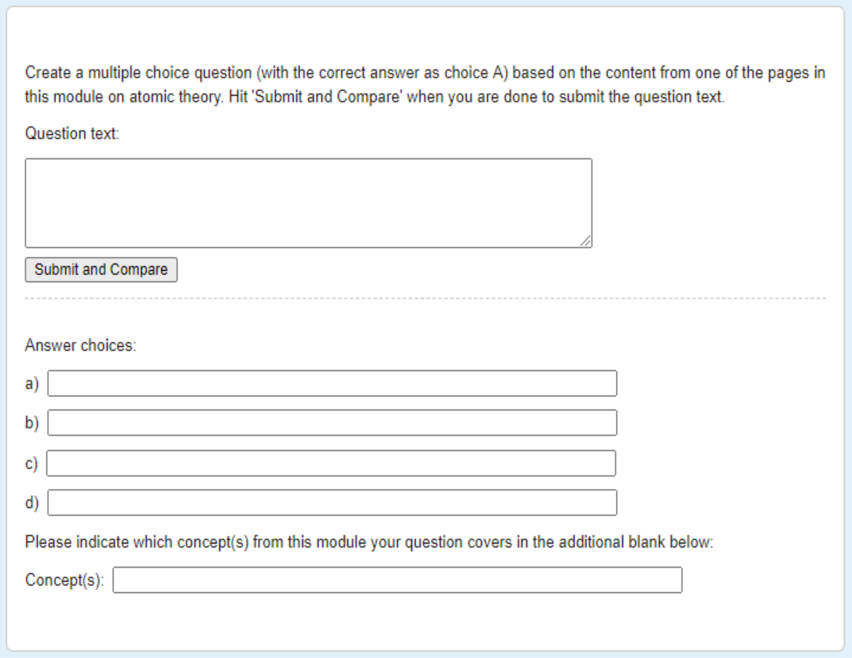
Course	Semester	Instructor	Student Count
chem 1a	summer	t1	47
chem 1b	fall	t1	55
chem 1c	fall	t1	27
chem 1d	fall	t2	23
chem 1e	fall	t3	2
chem 1f	fall	t3	23
chem 1g	fall	t4	24

Table 3.1: The seven introductory chemistry courses used in this study

Despite the offerings of the course having different instructors and even being used across different semesters, the students were provided with the same set of instructions regarding the use of the OLI materials. Students were provided with access to the OLI content, which served as supplementary materials for them alongside other course materials. They were not required to answer the questions found throughout the OLI modules or even access them. Students across all instances were granted access to the OLI content within the first two weeks of their respective course. They were also provided with an "Introduction to OLI" module, which provided an overview of how to effectively make use of the system and the concepts that will be covered in the course. All the instructional materials in OLI were optional to the students; there was no requirement for them to access or complete the materials. However, students were assessed on the concepts covered by the OLI materials, so it was beneficial for the students to utilize them.

The OLI content the students used for this study covers the topic of atomic theory and consists of six separate modules. Each module consists of several topic headers, containing paragraphs of instructional text and low-stakes activities embedded throughout. This particular section of the course consists of two learning objectives, where each module of the OLI content targets one of the two learning objectives. There are a total of 13 low-stakes and completely optional activities embedded throughout the six modules of the course, not including the activity used for this study. These activities include multiple choice questions, selecting the correct option from a dropdown, drag-and-drop exercises, and submitting a short answer to compare against an expert response. Each of these activities is broken down into steps, depending on the components of the activity, for a total of 37 unique steps. Every activity and their steps in the course provide students with feedback after they have been answered. Additionally, students have unlimited attempts to answer these questions, so they can continue until they are correct or choose to advance, regardless of a correct or incorrect response.

We focus on an activity we added to this course that involves the students creating a multiple-choice question, shown in Figure 3.1. This activity is found on the last module of the OLI content for this section of the course. This module provides several paragraphs of text that summarizes the content found on the five prior modules, along with this single activity. The activity is presented in the same low-stakes and optional format as the other 13 activities found prior in the course. It prompts students to create a multiple-choice question that targets content from one of the five other modules found in the OLI content. The students input the text for the question and then the correct answer, choice a, along with three distractors, choices b, c and d. Finally, they are asked to specify which specific concept(s) their question targets. We prompt them for the concept to help them focus their question on a specific topic found in the OLI content, rather than a broad and general chemistry question. Aside from that, no training or scaffolding was provided to the students to help them generate a question. We intentionally wanted to keep this low-stakes and optional, to examine the students' participation with the task and the quality of their contribution.


Create a multiple choice question (with the correct answer as choice A) based on the content from one of the pages in this module on atomic theory. Hit 'Submit and Compare' when you are done to submit the question text.

Please indicate which concept(s) from this module your question covers in the additional blank below:

Figure 3.1: The MCQ generation activity presented to students

3.2.2 Dataset

Student data was collected from their interactions with the 14 activities found in the course, including the MCQ generation task. However, since the MCQ generation task is our outcome, we focus our analysis on the 13 other activities that the students completed in the course, which consisted of a total of 37 unique steps. On average, an activity in the course consists of 3 unique steps, such as a single activity having the student select from three different dropdown menus. All of the activities found in the OLI course were completely optional, students could do as much or as little as they desired. For instance, sometimes a student would begin working on an activity, but not complete all of the steps. As a result, the system logs them having worked on that activity and also provides the exact number of steps for that problem that they completed. For this data set in particular, it is common for students to fully complete an activity if they start it, i.e., they will do all of the steps.

The total time students spend on solving activities in the course is also recorded by logging when the student first interacts with a step that is part of an activity, such as by clicking on it, and ending when they have made a submission for that step. This allows us to total the amount of time spent on the steps of an activity and calculate the total time a student spent on a given activity, which we can combine to get the total time spent on all activities in the course. In addition to these metrics of student participation and time spent, we have three metrics related to student performance on the activities. When a student works on a step for a given activity, OLI records if their first attempt at that step was correct or not. A first attempt at a problem can be a strong indicator of a student's current understanding of the concepts being assessed [48]. Relatedly, the total number of incorrect attempts made at a given step and the total number of correct attempts is recorded. These numbers can potentially exceed the total step count, as a student could correctly answer a question, then select an incorrect answer to see the feedback, then select the correct response once again, registering two correct and one incorrect for that step.

3.2.3 Calculating Question Quality

In order to assess the quality of the student-generated multiple-choice questions, we utilized a series of guidelines for identifying item-writing flaws (IWFs) in MCQs. The guidelines come from previous work that developed a taxonomy of 31 validated multiple-choice item-writing guidelines [86]. The exact rubric we used for the study was a modified version that consists of 19 unique items and has been used and validated in previous studies [31, 53, 181, 221, 222]. A full description of the 19 items that make up the rubric can be found in Table 3.2. In addition to the IWFs as a measure of question quality, we reviewed the cognitive level of each student-generated MCQ. Two levels of cognition were identified, recall or application, based upon a modified Bloom's Revised Taxonomy that MCQs have been evaluated under in previous studies [112, 151, 196, 199, 221]. A recall question, denoted by K1, assesses only the recall of facts or basic levels of comprehension. An application question, denoted by K2, assesses the higher level of cognitive ability focusing on application and analysis of the learned concepts.

Item-writing flaw	Definition
Ambiguous or unclear information	Questions and all options should be written in clear, unambiguous language
Implausible distracters	Make all distractors plausible as good items depend on having effective distractors
Use of none of the above	Avoid none of the above as it only really measures students ability to detect incorrect answers
Longest option is correct	Often the correct option is longer and includes more detailed information, which clues students to this option
Gratuitous information in stem	Avoid unnecessary information in the stem that is not required to answer the question
True/false question	The options should not be a series of true/false statements.

Convergence cues	Avoid convergence cues in options where there are different combinations of multiple components to the answer
Logical cues in stem	Avoid clues in the stem and the correct option that can help the test-wise student to identify the correct option
Use of all of the above	Avoid all of the above options as students can guess correct responses based on partial information
Fill-in-blank	Avoid omitting words in the middle of the stem that students must insert from the options provided
Absolute terms (never, always)	Avoid the use of absolute terms (e.g. never, always, all) in the options as students are aware that they are almost always false
Word repeats in stem and correct answer	Avoid similarly worded stems and correct responses or words repeated in the stem and correct response
Unfocused stem	The stem should present a clear and focused question that can be understood and answered without looking at the options
Complex or K-type	Avoid questions that have a range of correct responses, that ask students to select from a number of possible combinations of the responses
Grammatical cues in stem	All options should be grammatically consistent with the stem and should be parallel in style and form
Lost sequence in presentation of data	All options should be arranged in chronological or numerical order
Vague terms (sometimes, frequently)	Avoid the use of vague terms (e.g. frequently, occasionally) in the options as there is seldom agreement on their actual meaning
More than one or no correct answer	In single best-answer form, questions should have 1, and only 1, best answer
Negative worded stem (not, incorrect, except)	Negatively worded stems are less likely to measure important learning outcomes and can confuse students

Table 3.2: The rubric of 19 item-writing flaws used to evaluate the student-generated multiple-choice questions

Table 3.3 contains two different student-generated MCQs; the top question contains no IWFs and is at the application (K2) cognitive level. This question has zero IWFs according to the 19-item guideline, i.e. the question text is appropriately worded and all answer choices are plausible. It is at the application level of cognition as it requires the answerer to make a series of computations in addition to recalling multiple chemical elements, their atomic mass units (amu), and various counts of subatomic particles. In contrast, the bottom question contains two IWFs and is at the recall (K1) cognitive level. The first flaw is the logical cue in the stem, as it places an emphasis on the “neutral or uncharged” part, signaling that the correct answer is “neutrons” which could be guessed based on the similarity of the words alone. A second IWF occurs in the distractor choice of option d, “atom”, which is implausible due to the question stating that the particle is “in the atom”. Finally, this question is at the recall level of cognition because it is asking for part of the description of a neutron, which can be answered by simply recalling the definition of a neutron without any required application or analysis.

Three item raters evaluated each student-generated MCQ, following the 19 IWF guidelines. All three of the raters had content-area expertise, ample experience developing multiple-choice questions, and multiple prior training sessions in writing high quality assessments.

An unknown atom was found, tests have concluded that it weighed about 55 amu, and 29 neutrons were discovered. What element is the atom?	
a) Iron	b) Copper
c) Cobalt	d) Manganese
Which of these subatomic particles are neutral or uncharged in the atom?	
a) neutrons	b) electrons
c) electrons	d) atom

Table 3.3: A student-generated MCQ (top) that is K2 with 0 IWFs and another (bottom) that is K1 with 2 IWFs

Using the IWF rubric, the raters went through each of the 57 student-generated MCQs and applied the rubric to the question text and accompanying answer choices for each student contribution. While reviewing for IWFs, the raters also assigned a cognitive level of K1 or K2 to each question, based on if it required recall (K1) or application (K2) in order to answer the question. Although infrequent, three discordant questions were identified among the raters, related to multiple IWFs found in a single question. These discordant MCQs were discussed among the three raters until they reached a consensus on the categorization of IWFs for the three questions. Upon completion of the evaluation, all 57 student-generated MCQs were labeled with the count, if any, of IWFs they have and the cognitive level (K1 or K2) they assess.

3.2.4 Data Analysis

After the student-generated MCQs were evaluated for the IWFs and cognitive level to determine their quality, we began to analyze how the student interactions in the course correlated with both student participation on the task and the quality of their contribution. First, we investigated the different patterns of student participation in the course by looking at their interactions with the varying low-stakes activities and their steps embedded throughout the course. We also ran several unpaired t-tests to determine any significant differences between a student's interactions with the OLI materials and their participation with the MCQ generation task. Second, we use measures of central tendency to report the varying IWFs and cognitive levels of the student-generated MCQs. We also include a MannWhitney U-test for determining if there is a significant difference for students that generated K2 questions instead of K1. Third, we use a series of unpaired t tests to see which features of student behavior may lead to a higher quality contribution. Note that across all of the research questions there was no significant effect found based on the semester or instructor that the student had for the course. Additionally, a Bonferroni correction was applied to post-hoc tests used in the analyses that follow [15].

3.3 Results

3.3.1 Student Participation

Across all seven introductory chemistry courses used in this study, a total of 201 students accessed the OLI course. Among those 201 students, 57 of them completed the optional MCQ generation task. The course consists of a total of 14 optional low-stakes activities, including the MCQ generation one, and on average the students completed 9.75 of the 14 (69.94%) activities. Note that of the 201 students that accessed the course, 37 (18.41%) of the students did not interact with any of the 14 low-stakes optional activities found throughout the course.

To determine which features of student interaction in the course were indicative of their participation in the MCQ generation activity, we performed a series of t-tests on their behavior with the other activities found in the course. This revealed a significant difference between the student participation with the other low-stakes activities in the course and their participation in the MCQ generation task. An unpaired t-test showed there was a strong significant difference in the number of activity steps completed by students that did the MCQ generation tasks ($M = 45.24$, $SD = 4.22$) and those that did not do the task ($M = 24.39$, $SD = 19.49$), $t(199) = 7.917$, $p < .0001$. As expected, students that often completed all of the steps present in the activities embedded throughout the course were also more likely to also do the MCQ generation task. Similar significant results were observed for the number of activities done by a student and their participation for the MCQ generation task, t(199) = 7.087, p < .0001. This result supports the previous one, as the activities found throughout the course are composed of multiple steps and a subset of students completed all the 14 activities.

Due to student participation with all the activities and their steps being an indicator of their participation for the MCQ generation task, we also looked at the total time spent by the students on activities. On average, students spent roughly 18.5 (SD = 22.89) minutes working on the low-stakes activities found throughout the course. This was the time they spent interacting and answering the activities, which does not include the time they spent reading the instructional text and content. Students that did the MCQ generation task spent M1= 32.29 (SD_1 = 27.17) minutes working on other activities in the course while students that did not do the task spent an average of M2 = 13.03 (SD_2 = 18.47) minutes. There was a significant difference in the amount of time spent on activities between students who participated in the MCQ generation task and those who did not, t(199) = 5.787, p < .0001. This means students that answered most or all of the activities, thus spending more time on them, also participated more in the MCQ generation task.

3.3.2 Question Quality

To assess the quality of student-generated MCQs, we evaluated all 57 of their contributions using the 19 item-writing flaws rubric. This evaluation revealed a majority of the MCQs were of acceptable quality, with 22 (38.60%) containing no IWFs and 16 (28.07%) of the questions containing just one IWF. Table 3.4 shows the further breakdown of IWFs for all 57 MCQs that were evaluated, with roughly 33% of the questions containing more than one flaw. None of the contributions had more than four IWFs, and MCQs with one or fewer IWFs can be considered acceptable for use as a low-stakes assessment [221]. A total of 60 violations from 15 of the 19 IWFs were identified across the student-generated MCQs. While we utilized a 19-item rubric for the evaluation, only 15 of the criteria were present in the questions, as shown in Table 3.5. The four items that were not applicable to any of the MCQs were: negative word stem (not, incorrect, expect), more than one or no correct answer, vague terms (sometimes, frequently), and lost sequence in presentation of data.

Number of flaws	n (%) N = 57
None	22 (38.60%)
One	16 (28.07%)

Two	15 (26.31%)
Three	2 (3.51%)
Four	2 (3.51%)

Table 3.4: Total number of item-writing flaws encountered in the reviewed student-generated multiple choice questions

Item-writing flaw	n (%) N = 57
Ambiguous or unclear information	13 (21.67)
Implausible distracters	12 (20.00)
Use of none of the above	8 (13.33)
Longest option is correct	6 (10.00)
Gratuitous information in stem	3 (5.00)
True/false question	3 (5.00)
Convergence cues	3 (5.00)
Logical cues in stem	2 (3.33)
Use of all of the above	2 (3.33)
Fill-in-blank	2 (3.33)
Absolute terms (never, always)	2 (3.33)
Word repeats in stem and correct answer	1 (1.67)
Unfocused stem	1 (1.67)
Complex or K-type	1 (1.67)
Grammatical cues in sentence completion	1 (1.67)

Table 3.5: Frequency of item-writing flaws identified in the student-generated multiple choice questions

In addition to evaluating the MCQs based on the 19 item writing guidelines, we assessed the cognitive level of the 57 student-generated MCQs to further determine their quality. A vast majority of the questions (n = 49, 85.96%) were written at the K1 level, indicating that they focused on recall and comprehension. Interestingly, as shown in Table 3.6, the eight questions written at the K2 level of application and analysis had one or fewer errors. Although at a much smaller sample size, there was a significantly higher chance that a K2 question would have zero or one IWFs compared to a K1 question, as indicated by a Mann-Whitney U-test, U = 297, z = 2.308, p = 0.014.

Item-writing Flaws (%)	K1 - Recall & Comprehension	K2 - Application & Analysis
None	16 (28.07)	6 (10.53)
One	14 (24.57)	2 (3.51)
Two	15 (26.32)	0
Three	2 (3.51)	0
Four	2 (3.51)	0

Table 3.6: Cognitive level assessed by the student-generated question and the number of item-writing flaws it has

To determine which of the 57 student-generated MCQs were of acceptable quality, we grouped them into two categories based on their number of item-writing flaws. In total 19 (33.33%) of the questions were evaluated as being not acceptable, due to having two or more IWFs. The remaining 38 (66.67%) questions had either zero or one IWFs and were deemed to be acceptable for use. Table 3.7 presents an example of two student MCQs evaluated as acceptable, as they both have 0 IWFs and could be utilized as formative assessments in the course. Table 3.8 shows two student MCQs evaluated as unacceptable. In particular, the question on the top has unclear wording in the question's text, "number represent of the element", and has "none of the above" as an answer choice. The question on the bottom of Table 3.8 contains implausible distractors (i.e. Aristotle, who is never mentioned in the course) and has the longest and most detailed option as the correct answer.

An atom has an atomic number of 5 and a mass number of 11. How many neutrons are in this atom?

a) 5 b) 6

c) 16 d) 11

Which scientist discovered that protons are centered in the nucleus of an atom?

a) Rutherfod b) Thomson

c) Chadwick d) Milikan

Table 3.7: Two student generated MCQs evaluated by experts as being acceptable for use

What does the atomic number represent of the element?	
a) proton	b) neutron
c) electron	d) none of the above

Which physicist discovered the Cathode Ray Experiment?	
a) JJ. Thompson	b) Milikan
c) Aristotle	d) Leucippus

Table 3.8: Two student generated MCQs evaluated by experts as being unacceptable due to their Item-Writing Flaws

3.3.3 Student Interaction and Question Quality

We investigated if particular student interactions with the other low-stakes activities in the course correlated with the quality of their contribution, in order to see how we might predict or promote better questions from the students. While there was a significant difference found between student participation in the MCQ generation task and participation in the other low-stakes activities throughout the course, it was not found to correlate with the quality of the student contribution in this study (t(199) = 4.891, p = 0.417). However, student performance on the activities had a significant effect on the quality, measured in IWFs, of their MCQ contribution t(55) = 2.973, p < .005, as students who made more incorrect answers were more likely to contribute questions evaluated as unacceptable. There was a significant difference between students answering an activity correctly on the first try and the quality of their contribution, t(55) = 2.300, p < .05.

The previous findings relate to the student potentially having a better understanding of the material, thus making fewer mistakes and answering the questions correctly. This better understanding might in turn help the student to provide a higher quality question. In addition to student knowledge, we investigated if more time spent on the MCQ generation task led to a potentially higher quality contribution. However, the total amount of time a student spent on MCQ generation task (Mseconds = 153) and the quality of the contribution was found to not be statistically significant, t(55) = 0.4769, p = 0.6353).

3.4 Discussion

In this research, we investigated the effects of student participation and performance on their contribution to a MCQ generation task. We found that the students who chose to participate in the task generally completed all of the other optional activities found in the course. Even with the task being optional and only providing brief instructions with no scaffolding, students were able to generate MCQs that could be utilized as formative assessments for the course without any modifications. In exploring what features of student interaction in the course impacted the quality of the MCQ they generated, we found that their performance on the other low-stakes activities was significantly correlated with it. These findings suggest that students can create high quality multiple-choice questions from an optional and low-stakes activity within an online learning environment.

With all the low-stakes activities embedded throughout the course being completely optional, including the MCQ generation one, there was still a high amount of overall participation from the students. This was particularly surprising for a learnersourcing activity, which generally has lower participation rates due to the lesser perceived value students see in completion of the activity [77]. While past MCQ generation methods have relied on external systems [58, 109, 192, 233] or embedding the task in a high-stakes required assessment such as an exam [92], our study presented the task as a low-stakes activity, seemingly fitting in among the MCQ and drag-and-drop activities found on the other modules of course content. Leveraging just the native features of the system, in this case textboxes for short answer questions, we were able to provide students with the MCQ generation task seamlessly and without requiring them to utilize yet another platform. It is likely participation would be even greater if the task was required by students or embedded into a high-stakes assessment, such as a quiz question. However, this approach would introduce another series of potential complications, such as requiring it to be graded and potentially introducing an abundance of unacceptable questions contributed by students that do not wish to do the activity, but are forced to in the context.

Intuitively, student participation in terms of their interaction with the other low-stakes activities found in the course was strongly associated with their participation in the MCQ generation task. Since the course was relatively small, consisting of just 6 modules and 14 activities, it was common for the students to either complete all of the activities or choose to ignore them altogether. While we could not accurately calculate the exact time a student spent in the course, due to them potentially leaving the resources up while they are doing other work on the computer, estimates based on their access time and the time they spent on activities suggest the material took the students about two to three hours. Almost 20% of the students that accessed the course materials did not complete any of the activities in the course. One reason this might be the case is that they already had prior knowledge of the materials for this particular section of content, so they did not feel the need to do them. If this was the case, then we would want to also include those students in the MCQ generation task in order to take advantage of their existing knowledge. Encouraging all students to participate in this activity, such as through including a motivational prompt about how it can benefit their learning, could potentially better engage a student that might otherwise skip it.

Evaluation of the 57 student-generated MCQs identified that a majority of them had zero or one IWFs, with a very few number of questions having more than two IWFs. This came as a surprise considering that we tried to keep the activity brief and accessible with a concise instructional prompt for the task, no prior training being offered, and a lack of scaffolding being provided to the students as they worked through the task. It is possible that prior to the course, some students had experience writing MCQs or that they were particularly thoughtful and engaged with the activity since it was asking something non-traditional of them. However, this study demonstrates that even without training the students, providing them with overly detailed instructions, or even giving them MCQ writing guidelines, they can still contribute acceptable questions. However, the quality of the generated MCQ may be further improved by providing these resources to the students, but there are potential tradeoffs to consider between the brevity of the activity and the student participation garnered.

While the cognitive levels of the questions were mostly at the K1 level of recall and comprehension, this is typical of MCQs due to the nature of the assessment and is in line with findings from previous work [31, 221, 240]. Additionally, several student-generated MCQs did reach the K2 level of application and analysis and had significantly lower IWFs at this cognitive level. Further investigation remains on how we can better assist students in generating MCQs that target this K2 level, but MCQs at the K1 level are still usable for both formative and summative assessments.

A majority of the IWFs encountered in the student-generated MCQs presented themselves in the form of ambiguous or unclear information, which relates to the question stem being unclear. This flaw could be alleviated by providing the students with guidelines for question writing or reminding them to read over question text for clarity before submitting it. The second and third most occurring IWFs both relate to the answer choices of the question, as distractors are notoriously difficult to construct for MCQ generation [123]. Introducing a form of scaffolding to the activity which prompts students to think about the distractors they create or what constitutes an acceptable distractor, could potentially help students overcome these two common flaws. Ultimately these three most common IWFs that were identified in the contributions are not surprising, as they match findings from previous studies that reviewed MCQs generated by instructors [53, 222]. Interestingly, these studies suggest that no matter the expertise level, instructor or student, generating a quality MCQ free of flaws may still pose a challenge.

Grouping the student-generated MCQs by their potential IWFs resulted in roughly 33% of the questions being evaluated as unacceptable in their current state. Based on the expert evaluation, only a few of the questions were beyond repair. A majority of the questions that contained multiple IWFs could be resolved with a few minor edits. The question's central idea and what it is trying to assess was typically conveyed even with IWFs present, which allows for the question to potentially be leveraged later by another person to make corrections to it, akin to previous work for learnersourcing MCQs [109]. The other 67% of the questions were evaluated as being acceptable for use and could be directly utilized as formative assessments for the course in their existing state. Although these were acceptable, the cognitive level they assess could be enhanced from the K1 level, however, for a quick and low-cost way to assess student knowledge, they suffice.

There was no significant correlation between student participation in the low-stakes activities with the quality of the MCQ a student contributed. We expected that increased participation with the other activities would correlate with an improved quality of question, but there may have been a ceiling effect, since a majority of the students that did the MCQ generation task did most, if not all, of the other lowstakes activities. Features relating to student performance on the low-stakes activities embedded throughout the course, such as their first attempt correct or the number of incorrect answers they gave, were both significantly related to the quality of the student question. This finding is relatively intuitive in that students who make fewer mistakes overall and get the problem correct on the first attempt may demonstrate a higher level of mastery and thus can create higher quality questions using that knowledge. However, students who do not demonstrate mastery and are more novice bring about a unique perspective in the MCQ generation process and therefore they should not be overlooked. One way to effectively leverage this could be to have higher performing students generate the questions and then other students verify or improve it, akin to previous learnersourcing work [150]. For instance, potential student misconceptions may arise in the question text or answers they generate, which of itself are valuable insights into the student learning process that could potentially be leveraged as the source material for a question.

In our study, students chose to participate in a learnersourcing task with minimal instructions where they generate a MCQ, even when it is presented as just another low-stakes optional activity. A new system or excessive information does not necessarily need to be introduced to the students to have a successfully generated MCQ. Keeping it native and simple worked surprisingly well in this study, compared to the participation rates detailed in previous learnersourcing studies [123, 227]. A majority of the student contributed MCQs could be utilized in their current state as formative assessments in the course, since they contained zero or just a single IWF that was not a cause for rejection. Even the contributed questions that contained multiple flaws could potentially be remedied with just a few edits to the question wording, although domain knowledge is often required for modifications to the distractors. While a majority of the questions assessed knowledge at the K1 level, that is a typical level for MCQs and sufficient for the amount of time and lack of training the students had [207, 240]. Providing question writing guidelines for students could potentially help them construct MCQs at a more advanced cognitive level.

3.5 Limitations & Future Work

There are several limitations in this study. First, our study relies on student data from optional activities found throughout the course. Our findings are prone to a self selection bias, as participants in the MCQ generation task might be the most driven students that want to complete all of the materials. Second, the students in this study were all enrolled in a summer or fall offering of the same introductory chemistry course, taught by one of four instructors. Thus, our results may not generalize as well to other domains with students that are pursuing different degrees and coursework. Finally, our sample size could be increased to gain more statistical significance and insight into the cognitive level of the MCQs generated by students. With only 8 of the 57 student-generated questions targeting the K2 level of application and analysis, increasing the sample size could yield a more accurate measure of ratio. However, our current sample size is large enough to detect statistical significance.

In this work, our analysis is limited to the students' performance within the context of the unit in the OLI system. Ideally, we would like to include other summative measures of the student and their learning, such as their grade in the course or GPA. We urge future research to investigate how we can incorporate learnersourcing tasks, such as MCQ generation, through more low-stakes and natural mechanisms. While having students complete such tasks as part of a quiz or homework assignment may yield greater participation, it may not be the best use of required student time, especially if it requires them to become familiar with an entirely new system. Investigating the learning differences between students completing such a task in a low-stakes vs. high-stakes environment may yield interesting results for future tasks. In this study, we intentionally had the task be low-stakes and use minimal instruction. We plan to investigate the trade-off between instructional brevity and student participation in learnersourcing activities in the future.

Additionally, while these results might suggest higher performing students generate higher quality questions, we should not overlook the other students. Every student has a valid viewpoint they can bring into the MCQ generation process, not just the top performing ones. For instance, leveraging the full range of students can yield questions that target misconceptions that might otherwise be overlooked. Future work may look to analyze whether MCQs generated by lower performing students reflect potential misconceptions they hold, as a way to both identify and remedy potential knowledge gaps. Finally, towards building up a practical question bank for instructors, we will look into diversifying the topics of generated questions, such as placing the activity at different points throughout the course.

3.6 Conclusion

This work demonstrates that students' participation and performance with activities in an online chemistry course correlates with their contribution to a MCQ generation task. Our results highlight how student behaviors regarding their completion of the low-stakes activities in the course are indicative of their participation on an optional learnersourcing task. Requiring the student to do the activity, such as putting it in a high-stakes assessment, having students use another system, or utilizing another tool to generate the MCQ, is not necessarily required for contributions that are evaluated as quality questions. Students are capable of providing recall and comprehension level MCQs, without detailed instructions, prior training, or scaffolding. While a majority of the MCQs students generated were acceptable and could be used as is, there is an influence of student performance, in terms of making fewer incorrect answers and getting questions correct on the first attempt, on an improved quality question. This research helps demonstrate one way to help scale online learning and improve educational resources, by leveraging the students in a course. Not only can these created questions aid the instructor and other students, but the process of students generating these questions has been shown to benefit their learning. This work opens up further opportunities for both engaging students in the process of generating MCQs and promoting their behavior that leads to a higher quality contribution for future learnersourcing tasks.

Chapter 4
Students Generating High Quality SAQs

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. "Assessing the quality of student-generated short answer questions using GPT-3." In *European conference on technology enhanced learning*, pp. 243-257. Cham: Springer International Publishing, 2022.

4.1 Introduction

Students generating short answer questions has been proven to support their learning of new instructional content [22, 43]. As students generate questions, they deeply engage with the subject matter and utilize critical thinking skills [57]. This process leverages student engagement in ways that provide meaningful data around student interaction integrated with new student-generated learning assets that can support future learners [61]. This is known as a form of learnersourcing, where students complete activities that produce content which can then be leveraged by future learners [110]. Several systems to support students in the generation and sharing of questions have been leveraged by thousands of students [58, 109]. This usage has led to the student-authoring of nearly a million questions, while also supporting research demonstrating that student question generation can lead to positive learning outcomes [107].

On the other hand, the quality of student-generated questions can widely vary [162]. While existing learnersourcing tools can scaffold this process and guide students towards generating better questions, they often require external systems [58, 109]. Additionally, evaluating the multitude of student-generated questions presents another challenge, with past research relying on experts, other students, or automated methods [140]. Automated methods often rely on the surface-level features of the question, such as the readability of text length, without including the pedagogical value it adds to a course. Recent research has developed and utilized a rubric for human evaluation of automatically generated questions that includes both linguistic and pedagogical criteria [91, 217]. However, these criteria have not seen wide adoption in automated evaluation methods, largely due to the difficulties associated with encoding them in a machine-interpretable way.

In this work, we explored how students could contribute short answer questions with minimal scaffolding and how we could assess their quality using machine learning models that match expert evaluations. We deployed a short answer question generation activity into seven instances of an online college-level chemistry course. From the student responses, we evaluated the quality of the short answer questions, determining if they were of sufficient quality, with respect to their pedagogical value, to be used in the course. The student-generated questions were also assessed for their cognitive level, in terms of Bloom's Revised Taxonomy [119]. Following this, we explored automatically evaluating the questions for their quality and cognitive level using a state-of-the-art language model.

Our work makes the following contributions towards learnersourcing and question evaluation. First, we demonstrate that students can create high-quality questions with a simple prompt that can be added to virtually any learning platform. Second, we present an expert evaluation process investigating the quality and cognitive level of student generated questions. Third, we evaluate the usefulness of using a state-of-the-art language model in classifying educational questions, in an effort to make this process scalable and potentially saving instructor time. Ultimately, our work demonstrates how students can generate high quality questions with minimal scaffolding and how language models might be leveraged to assist in the quality and pedagogical evaluation of short answer questions.

4.2 Learning Platform and Data Collection

The present study takes place in a digital courseware platform known as the Open Learning Initiative (OLI). OLI is an open-ended learning environment that offers courses from a variety of domains and consists of interactive activities and diverse multimedia content [27]. OLI consists of instructional content and low-stakes, also known as formative, activities. These activities consist of a variety of question types such as multiple-choice questions, short answer, and dropdown style questions. Students work through different modules in the system, akin to chapters in a textbook, where they are presented with instructional text and videos. Low-stakes activities are embedded throughout these instructional materials, providing the students with feedback and practice opportunities to assess the concepts they are learning.

The data used in this study was collected from a week-long module in seven instances of an introductory chemistry course taught at a community college in the western U.S. The course consists of first- and second-year undergraduates from varying degree backgrounds, with most of the students pursuing a chemistry-related degree. The data comes from the fall semester of 2021, when the introductory chemistry course was offered in the OLI system. In total, the data consists of 143 students and their contribution to the short answer generation activity. The OLI content the students used during the week when our data was collected covers the topics of pH, buffers, and amino acids. There are a total of 38 low-stakes activities embedded throughout the pages of this module.

Every activity provides the students with detailed instructional feedback, for both incorrect and correct responses.

We focus on an activity that was added to the course that involves each student generating a short answer question. In the chemistry course, this activity is found on a page containing four paragraphs of instructional text, three worked examples, and eight multiple-choice questions. This activity is presented in the same low-stake format as the other activities found throughout the course, as students do not receive a grade for their participation or the quality of their response in the activity. It prompts students to generate a short answer question, by asking them to "*Create a short answer question that can be correctly answered based on the content covered in this module*". In the activity, students are first prompted to write the question text in the provided text box on the top part of the activity and then write the answer to the question in the bottom text box. The instructions for the self-explanation are intentionally brief and similar prompts have been used in related studies by [4, 239].

4.3 Data Analysis

4.3.1 Human Evaluation

The 143 student-generated short answer questions were evaluated by two experts to assess their quality and Bloom's Revised Taxonomy level. The two experts had content knowledge in chemistry, multiple years of teaching experience, familiarity with the OLI course, and ample previous experience coding qualitative student data. To first evaluate the quality of the questions, the two experts used a 9-item rubric that has been used in previous studies for assessing the linguistic and pedagogical quality of questions [91, 217]. This rubric contains 9 hierarchical criteria, shown in Table 4.1. These criteria are asked to the two experts in the order, from top to bottom, that they are presented in the table. Eight of the rubric criteria involve binary (yes/no) responses. The only non-binary item is *information needed*, which consists of three unique options, where each corresponds to the location of the information the students need to know in order to successfully answer the question.

The rubric items are hierarchical by nature, meaning that if certain criteria are answered as "no", then the remaining items will be marked as "not applicable". These criteria are bolded in Table 4.1. For example, if the experts answer "no" to the *answerable* rubric item, then the three items that follow will be marked as "not applicable". This contributes to avoiding distortion of the rubric criteria distributions for questions that are not ratable across certain items and helps to save the expert evaluators' time. The inter-rater reliability (IRR) values between the two evaluators for each rubric item are also reported in Table 4.1. It includes the percentage agreement and Cohen's Kappa κ statistic [148] as a measure of IRR for all rubric items. These items are at either a near perfect or substantial level of agreement between the two raters. Two of them, *domain related* and *central*, had perfect agreement, as all of the student-generated questions pertained to chemistry content covered in the current OLI module.

Rubric Item	Definition
Understandable (97.20%, κ = 0.83)	Could you understand what the question is asking?
DomainRelated (100%, κ = 1.0)	Is the question related to the Chemistry domain?
Grammatical (96.15%, κ = 0.82)	Is the question grammatically well formed, i.e. is it free of language errors?
Clear (98.46%, κ = 0.83)	Is it clear what the question asks for?
NotRephrasing (89.52%, κ = 0.66)	Does the question assess course content that has not been assessed by an existing question in the course?
Answerable (99.19%, κ = 0.88)	Are students probably able to answer the question?
InformationNeeded (88.14%, κ = 0.73)	(op) Information presented directly and in one place only in the text (dp) Information presented in different parts of the text (te) A combination of information from the text with external knowledge
Central (100%, κ = 1.00)	Do you think being able to answer the question is important to work on the topics covered by the current module?
WouldYouUseIt (82.35%, κ = 0.62)	If you were a teacher working with the OLI module in your class, would you include this question in the course?

Table 4.1: The hierarchical 9-item rubric used to evaluate the questions, the bolded criteria stop the review process if answered as “no”. The bracketed numbers indicate agreement percentage between raters and Cohen’s κ value for each item

If the expert evaluators answer “yes” to all the binary rubric items and answer any of the three options for *information needed* then we consider that to be a high quality question. In line with previous research, meeting all the rubric criteria suggests that the question is both linguistically and pedagogically sound [91, 217]. Additionally, the last rubric criteria *would you use it* asks the evaluators if they would use the student-generated question if they were teaching the course and using the OLI materials. As the evaluators are familiar with the OLI content and have prior teaching experience, they can judge the pedagogical quality of the student-generated questions. However, we acknowledge that despite the two expert evaluators’ backgrounds and high IRR they can still interpret the student-generated questions in different ways as influenced by their prior knowledge and linguistic preferences [11].

In order to assess the cognitive level of the student-generated questions, the two expert evaluators utilized Bloom's Revised Taxonomy [119]. This taxonomy, shown in Table 4.2, has been applied to educational questions in prior research [93, 242]. It consists of six different levels, where each one corresponds to the cognitive processes involved in answering the question. Using these six taxonomy levels, the two expert evaluators classified each student-generated question to a level, depending on what cognitive process is required to answer it. Note, only student-generated questions that had no "non applicable" answers to the nine rubric criteria were evaluated in this way, resulting in a total of 120 of the 143 (84%) questions being assigned one of the six levels as agreed upon by the two expert evaluators. While there are six levels to the taxonomy, the student generated questions in this study were all assigned to the first four levels, as none of the questions targeted the cognitive processes of *evaluate* or *create*. The omission of these two levels was not by design, however they are less common for short answer questions typically found in courses, which are more likely to assess the first four levels of Bloom's Revised Taxonomy [209]. Additionally, while assessing the questions using the 9-item rubric and for Bloom's Revised Taxonomy, the two expert evaluators had disagreements, as indicated by the Kappa values in Table 4.1. The discordant criteria for such questions were discussed between the two raters, resulting in them reaching a consensus on the categorization of the question.

Bloom's Level	Definition
Remember	Retrieve relevant knowledge from long-term memory
Understand	Construct meaning from instructional messages, including oral, written and graphic communication
Apply	Carry out or using a procedure in a given situation
Analyze	Break down the learning material into constituent parts and determine how parts relate to one another and to an overall structure
Evaluate	Make judgments based on criteria and standards
Create	Put elements together to form a coherent whole or to reorganize into a new pattern or structure

Table 4.2: Six levels of Bloom's Revised Taxonomy [119] in ascending cognitive order from lowest to highest, along with their operational definitions

The IRR between the two expert evaluators for applying Bloom's Revised Taxonomy to the student-generated questions was assessed via percentage of agreement (81.67%) and Cohen's Kappa (κ=.74), suggesting a substantial level of agreement. This agreement level is akin to previous studies that applied Bloom's Revised Taxonomy to student generated questions [238]. In accordance with previous research [119, 235], we define a student-generated question as assessing a low cognitive level if it was evaluated to be at the *remember* or *understand* levels. Conversely the question is said to assess at a high cognitive level if it was evaluated to be at the *apply*, *analyze*, *evaluate*, or *create* levels. Typically, multiple-choice and short answer questions rely on the cognitive processes associated with lower cognitive levels, although both question types can assess higher levels [229]. It is desirable to have questions assessed at a higher level, as it is more beneficial for student learning [119].

4.3.2 Human Evaluation

Our second evaluation method utilizes GPT-3, a language model with up to 175 billion parameters trained on a large dataset of text scraped from the internet [33]. We selected this language model for our evaluation due to it being state-of-the-art for multiple natural language processing tasks and being the largest publicly available transformer language model. It is a high-performing and popular language model choice for text classification, with recent applications in classifying emails [223] and determining if news articles were real or fake [39]. In this work, we used GPT-3 to perform classification on the student generated questions in two different ways. We avoided using typical automated question generation evaluation criteria such as BLEU or METEOR, as they have been proven to not correlate with human evaluation and do not have pedagogical implications [206].

First, we used it for binary classification to see if it could classify the student generated questions as being low or high quality, matching the evaluation of the two experts. To make this classification, we first fine-tuned a GPT-3 Ada model on the LearningQ dataset [41], which is publicly available and contains 5,600 student-generated short answer questions from Khan Academy. Each question in this dataset was evaluated by two expert instructors and assigned a label corresponding to if it was *useful for learning or not*. The researchers for the LearningQ dataset defined a question as being *useful for learning* akin to several of the rubric criteria we utilized in this study. They based their evaluation on the following three criteria: (i) concept-relevant, seeking information on the concepts taught in the course; (ii) context-complete, providing enough information to be answerable by other students; and (iii) not-generic, meaning the question asks about a course concept not on another topic or of another style, such as asking for learning advice. Additionally, the questions in the LearningQ dataset came from a variety of domains, which included STEM courses and a single humanity one. No preprocessing was performed on the questions used to fine–tune the model; they were used as-is from the publicly available dataset along with their corresponding binary labels. Fine-tuning the model with default hyperparameters suggested by the documentation[1] took approximately 10 min and incurred a cost

[1] We used the default hyperparameters as suggested in https://beta.openai.com/docs/guides/fine-tuning

of $0.21. Upon completion, we passed in the student-generated questions as the GPT-3 model's input, obtaining the output as a binary label indicating if it rated each question as useful for learning (*high quality*) or not (*low quality*).

Secondly, we used another instance of the GPT-3 Ada model to perform multiclass classification using Bloom's Revised Taxonomy levels. We once again use GPT-3 Ada, which was selected due to its low cost and effectiveness at classification tasks that are less nuanced, with comparable performance to the Davinci model. We wanted to determine if GPT-3, fine-tuned on example questions from each level, could perform similarly to the two expert evaluators. To fine-tune the model, we utilized a dataset consisting of 100 questions mapped to each of six Bloom's Revised Taxonomy levels, for a total of 600 questions [235]. These 600 questions were assigned to a level of Bloom's Revised Taxonomy by a pedagogical expert and this dataset has been used in ample previous studies involving fine-tuning and classification tasks. In the present student, the expert evaluation of the student-generated questions only identified four of the six Bloom's levels that were applicable to the questions. However, we included questions from the two unused Bloom's levels in the fine-tuning process, because if the model was accurate, we could utilize it for future datasets that may contain questions at that cognitive level. For this dataset, we performed no preprocessing on the questions used to fine–tune the model; they were used as-is from the publicly available dataset along with their corresponding Bloom's Revised Taxonomy labels. We once again fine-tuned the model with default hyperparameters which took approximately 5 min and incurred a cost of $0.08. Upon completion, the student-generated questions were passed as the GPT-3 model's input, outputting Bloom's labels for each question.

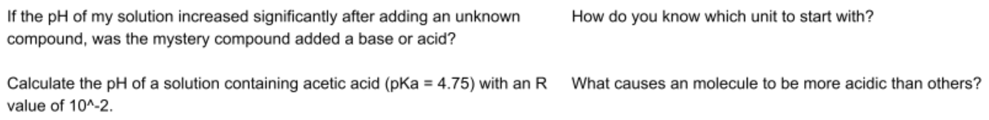
4.4 Results

We first begin with our human evaluation by experts, using the 9-item rubric, across all 143 student-generated short answer questions. As indicated in the Data Analysis section, the rubric criteria are hierarchical and they can be marked as "not applicable", causing the following rubric items to be ignored. For example, if a question was marked "not applicable" for the first rubric criteria of *understandable*, that would reduce the question pool for the other eight criteria. We report the percentage relative to the remaining questions, followed by the absolute percentage, i.e. (relative % / absolute %).

4.4.1 Short Answer Question Quality

We found that 91% of the student-generated short answer questions were rated *understandable*. All the questions rated as *understandable*, were also rated *domain related* (100%/91% total). Most questions were also free of *grammatical errors* (90%/82% total), which includes typos and punctuation mistakes. As a question’s clarity is related to the understandability of the question, there were also many questions (95%/87% total) that were evaluated as being *clear*. If a question assessed course content that has not been assessed by an existing question found somewhere in the module, then it was marked as *not rephrasing* (84%/73% total). This is one of the lowest rubric criteria percentages and also presented a challenge for the evaluators to find agreement on, as they achieved a Cohen’s Kappa of $\kappa = .66$.

The evaluation shows that most of the questions are rated as *answerable* by future students in the course (97%/84% total). Similar to the criteria about being domain related, the *central* criteria (100%/84% total) was perfect for the remaining pool of questions. This not only means the question relates to the chemistry, but it specifically targets a concept that is addressed in the current module. According to the evaluators, knowledge required for answering the questions is obtained in *one place* (68%/57% total) or in *different places* (30%/25% total) throughout the module. However, there were two questions that were evaluated as needing both the instructional *text and external knowledge* (2%/1% total).



If the pH of my solution increased significantly after adding an unknown compound, was the mystery compound added a base or acid?	How do you know which unit to start with?
Calculate the pH of a solution containing acetic acid (pKa = 4.75) with an R value of 10^-2.	What causes an molecule to be more acidic than others?

Figure 4.1: The two questions on the left are evaluated as being high quality and the two questions on the right are low-quality, due to being vague (top) and grammatically incorrect (bottom)

As described in the Data Analysis section, a question was categorized as high quality if it passed all nine rubric criteria, including being evaluated as *would you use it* (38%/32% total). In total, 46/143 (32%) student-generated short answer questions met this criterion by passing all nine rubric items and were deemed to be of high quality. Figure 4.1 shows two questions evaluated as high quality and two questions evaluated as low-quality. The question in the upper-right was evaluated as not being *understandable* and the question in the bottom-right was not *grammatical*.

4.4.2 Higher Order Cognitive Processes

In order to assess the cognitive-level of the student-generated questions, the evaluators applied Bloom’s Revised Taxonomy to them. Due to some of the questions having certain rubric criteria marked as “not applicable” and thus ending the review, 120/143 (84%) student-generated questions were assigned a Bloom’s Revised Taxonomy level by the evaluators. The majority categorization was remember (52%), with understand (25%) and apply (20%) being tagged to a similar number of questions, followed by analyze (3%). An example of the student-generated questions corresponding to each of these four Bloom's Revised Taxonomy levels is shown in Table 4.3.

Student-Generated Question	Bloom's Level
What is the point in a titration curve that indicates the pKa value of a weak acid?	Remember
Imagine an acidic solution with a low pH. If a strong base is added to the solution, what happens to the pH in relation to the pKa?	Understand
If 10mL of a diprotic weak acid is fully deprotonated with 20mL of 0.5M NaOH, how many moles of the acid and NaOH are there?	Apply
When stomach acid enters the esophagus, typically with a pH of 1.5 to 3.5, calcium carbonate is often used to combat this. Why would calcium carbonate be a good substance for this problem?	Analyze

Table 4.3: An example of a student-generated question assessed at each of the four levels of Bloom's Revised Taxonomy present in this study

Prior research [119, 209] has indicated that questions at the apply level and above are categorized as targeting higher order cognitive processes. As a result, 28/120 (23%) questions tagged with Bloom's Revised Taxonomy were evaluated as assessing at this higher level. Since Bloom's Revised Taxonomy level was not included in the criteria for a high-quality question, we investigated if there was a correlation between the two measures. Fisher's exact test revealed that there was a strong statistically significant association between the quality of the question and the cognitive level (p = .003). Figure 4.2 shows the distribution of Bloom's Revised Taxonomy levels between questions evaluated as being low and high quality.

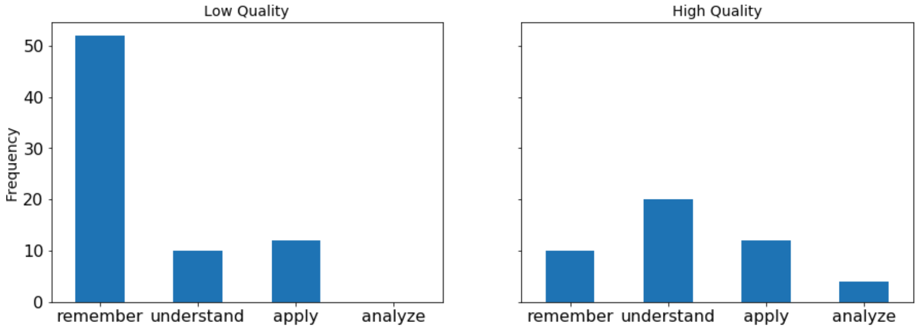


Figure 4.2: The distribution of the four Bloom's Revised Taxonomy levels between questions evaluated as low and high quality

4.4.3 Automatic Evaluation

We utilized the first fine-tuned GPT-3 model to classify the quality of the student-generated questions as either low or high quality. The model agreed with the human evaluation for 57/143 questions (40%). In the cases they disagreed with, 85/86 mismatches were interpreted as having high quality by GPT-3 but low quality by expert raters. There were only 13/143 questions (9%) the model classified as low quality, suggesting it was overestimating the quality of the questions, as 97/143 (68%) were evaluated by the experts as being low quality. Figure 4.3 provides a confusion matrix for the quality classifications made by the model.

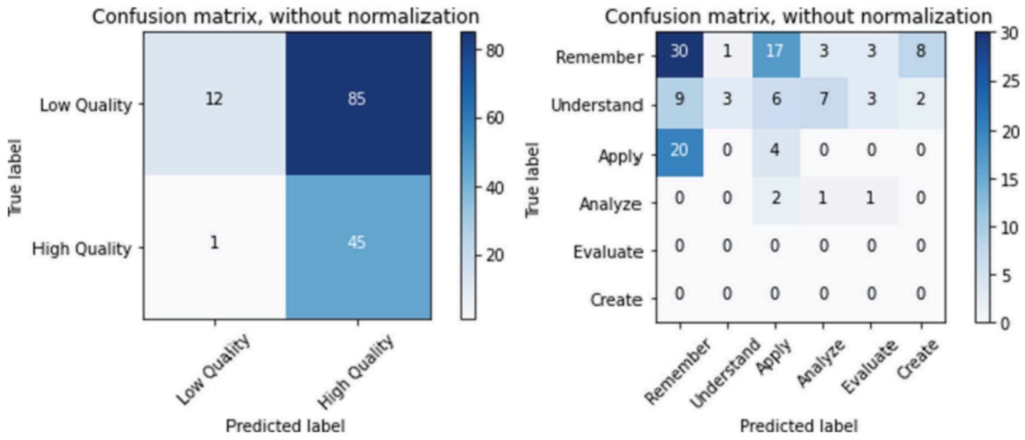


Figure 4.3: Confusion matrices for the classification of a question's quality (left) and Bloom's Revised Taxonomy (right)

We used the second fine-tuned GPT-3 model to classify the 120 student-generated questions to which the expert evaluators had assigned a Bloom's Revised Taxonomy level. The results of the model compared to the expert evaluation, including the percentage of matches for each Bloom's Revised Taxonomy level between the two, are shown in Table 4.4. In total, the model matched the expert evaluation for 38/120 (32%) student-generated questions. The GPT-3 model has a similar distribution of *remember* and *apply* questions, although they are often not correctly applied to the questions according to the expert evaluation. Additionally, GPT-3 classified 17 of the questions into the two highest cognitive levels that were not observed in our student-generated questions. Additionally, Figure 4.3 also provides a confusion matrix for the classification of Bloom's Revised Taxonomy between the expert human evaluators and the model.

Bloom's Level	Remember	Understand	Apply	Analyze	Evaluate	Create
Expert Evaluation	62	30	24	4	0	0
GPT-3	59	4	29	11	10	7

Matching %	48%	10%	4%	25%	0%	0%

Table 4.4: A breakdown of the six Bloom's Revised Taxonomy and the number of questions the experts and GPT-3 tagged to each level

4.5 Discussion

In this research, we utilized human experts and automatic methods to evaluate the quality and cognitive level of student-generated short answer questions. We found that students were able to contribute high quality questions, as evaluated by a 9-item rubric that contained criteria assessing the linguistic and pedagogical features of the questions. In total, 32% of the student-generated short answer questions were evaluated as being high quality, indicating that the evaluators could use them in the course in their present condition. Students generated these questions through a simplistic prompt consisting of a single sentence instruction and two textboxes embedded into a digital learning platform. Previous research often has an overall lower percentage of high-quality questions and utilizes external systems or scaffolding methods that require the students to spend more time on the question generation activity [4, 22]. We believe that the implementation we used in this study keeps students more engaged in the learning process, by allowing them to create the question in a more natural context as they work through the instructional text and assessments in the platform.

The cognitive processes that the student-generated questions target were evaluated by the two expert evaluators, which identified 23% of the questions as assessing at a high cognitive level and the remaining 77% assessing the lower two cognitive levels. This majority distribution of the short answer questions assessing at the *remembering* and *understanding* cognitive levels is in line with findings from previous work [10, 242]. These questions that assess the first two cognitive levels can still be effective, particularly when students are first learning new concepts, where they might need to first learn essential terminology, methods, and formulas [119].

Automatic evaluation of the student-generated questions for both their quality and cognitive level was suboptimal compared to previous work leveraging different language models [41, 195], however, such prior research often evaluates questions that are mostly at the remembering cognitive level and often involve basic reading comprehension with no domain-related knowledge being assessed, which are more appropriate for students at lower education levels [122]. The student-generated questions in this study were at the post-secondary education level, assessed chemistry knowledge, and often included domain terminology. These differences between questions used in prior research in this study likely contributed to the difficulty the two GPT-3 models had, even when they were fine-tuned on relevant data for the classification tasks. The percentage of expert matching classifications the models achieved for the quality (40%) and cognitive level (32%) could provide an initial estimation of the questions' value.

4.6 Limitations & Future Work

The main limitation of this study comes from the dataset, as the 143 student-generated short answer questions that were analyzed were all in the domain of chemistry. Including student-generated questions from other domains could lead to more generalizable findings. Question evaluation often entails human annotations as the ideal criterion to compare automatic methods against; however, there is a subjective nature to human ratings. While we tried to reduce subjectivity by using a detailed rubric for the human evaluation and achieving a high IRR for each criterion, there still lies the potential for different evaluation depending on who is doing the evaluation. Finally, the results of the GPT-3 model were suboptimal, often overestimating the quality of the student-generated questions or misclassifying Bloom's Revised Taxonomy level. The results of these classifications were influenced by the datasets used to fine-tune them, which was limited by public datasets that classify the educational quality of the question and the cognitive level.

4.7 Conclusion

This work demonstrates that students can generate short answer questions that are both linguistically and pedagogically sound without requiring an external tool or scaffolding. In total, we found that 32% of all the student-generated questions were evaluated as being high quality by the expert evaluators. Across all the questions that were classified according to Bloom's Revised Taxonomy, 23% were evaluated as assessing high cognitive levels. Our results highlight how students in the context of an online course can create short answer questions that can readily be implemented into the course, providing new assessment opportunities for essential concepts. While the automatic evaluation may be improved with more robust datasets for fine-tuning, it offers a sufficient first pass classification that may assist experts in their evaluation of the questions. This research helps demonstrate one way to help scale online learning and improve educational resources, by leveraging the students in a course. It opens further opportunities for engaging students in the process of question generation and leveraging both humans and language models to assist in the evaluation process.

Chapter 5
Crowdsourcing Skill Tags for Assessments

> This chapter is based upon the following two previously published works:
>
> Moore, Steven, Huy A. Nguyen, and John Stamper. "Evaluating crowdsourcing and topic modeling in generating knowledge components from explanations." In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*, pp. 398-410. Springer International Publishing, 2020.

5.1 Introduction

The combination of data-driven knowledge tracing methods and cognitive-based modeling has greatly enhanced the effectiveness of a wide range of educational technologies, such as intelligent tutoring systems and other online courseware. In particular, these systems often employ knowledge component modeling, which treats student knowledge as a set of interrelated KCs, where each KC is “an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks” [116]. Operationally, a KC model is defined as a mapping between each question item and a hypothesized set of associated KCs that represent the skills or knowledge needed to solve that item. This mapping is intended to capture the student’s underlying cognitive process and is vital to many core functionalities of educational software, enabling features such as adaptive feedback and hints [167].

While machine learning methodologies have been developed to assist in the automatic identification of new KCs, prior research has shown that human judgment remains critical in the interpretation of the improved model and acquisition of actionable insights [133, 173]. An emerging area that has the potential to provide the human resources needed for scaling KC modeling is crowdsourcing. Naturally, the challenge with this approach is that the population of crowdworkers is highly varied in their education level and domain knowledge proficiency. Therefore, as a first step towards examining and promoting the feasibility of crowdsourced KC modeling, we studied how crowdworkers can provide insights into different word problems that might suggest areas of improvements and generating KCs for the questions. We took these insights via explanations, coded them and ran them through two topic models to analyze how they might be utilized for the task.

5.2 Methods

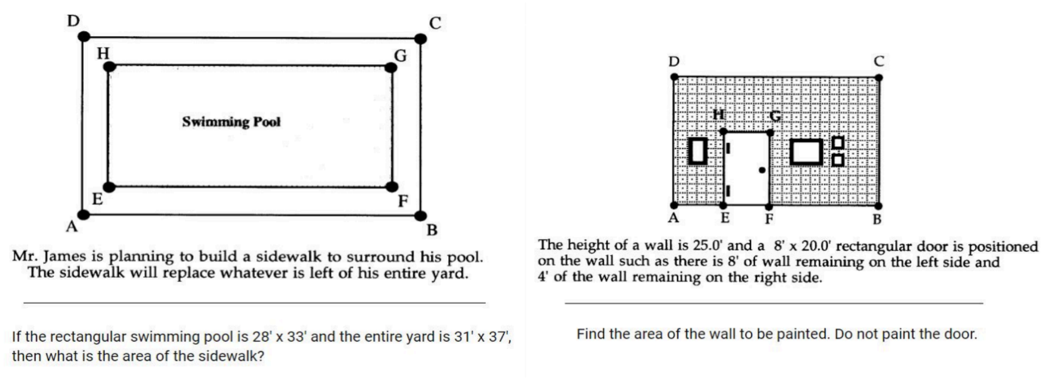
Our study consists of two experiments with the same procedure, but involve different domain knowledge. The first domain is mathematics, with a focus on the area of shapes; the second is English writing, with a focus on prose style involving agents and clause topics. In both domains, we deployed an experiment using Amazon's Mechanical Turk (AMT). Forty crowd workers on AMT, known as "turkers," completed the math experiment, and thirty turkers completed the writing experiment, for a total of 70 participants. In each domain, the tasks took roughly five minutes. Participants were compensated $0.75 upon completion, providing a mean hourly wage of $9.

The main task of the experiment presented participants with two word problems positioned side by side, labeled Question 1 and Question 2. In the math experiment, both problems involve finding the area of two different structures. In the writing experiment, both problems involve identifying the agents and actions of two different sentences. Participants were truthfully told that past students were tested on these problems and that the collected data indicates Question 2 is more difficult than Question 1. They were then asked to provide three explanations on why this is the case. The specific question prompt stated: "*Data shows that from the two questions displayed above, students have more difficulty answering Question 2 than Question 1. Please list three explanations on why Question 2 might be more difficult than Question 1*".

5.2.1 Math and Writing Experiments

The two mathematics word problems used for the explanation task can be seen in Figure 5.1. These problems come from a previous study of a geometry cognitive tutor [216], where the data indicates that students struggle more with the problem involving painting the wall (the right side of Figure 5.1). Both problems are tagged with the same three KCs by the domain experts that created the problems, so they assess the same content. These KCs are: Compose-by-addition, Subtract, and Rectangle-area.

Both problems used in the writing experiment come from an online prose style course for freshman and sophomore undergraduates (Figure 5.2). Similar to the math problems, student data collected from the online course indicates students struggle more with one problem over the other. The KCs were generated by domain experts and are: Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form.



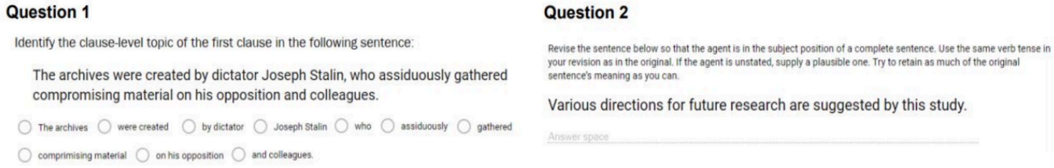
Mr. James is planning to build a sidewalk to surround his pool. The sidewalk will replace whatever is left of his entire yard.

If the rectangular swimming pool is 28' x 33' and the entire yard is 31' x 37', then what is the area of the sidewalk?

The height of a wall is 25.0' and a 8' x 20.0' rectangular door is positioned on the wall such as there is 8' of wall remaining on the left side and 4' of the wall remaining on the right side.

Find the area of the wall to be painted. Do not paint the door.

Figure 5.1: The two word problems for which participants provided three explanations in the math experiment, with the one on the right being more difficult



Question 1

Identify the clause-level topic of the first clause in the following sentence:

The archives were created by dictator Joseph Stalin, who assiduously gathered compromising material on his opposition and colleagues.

○ The archives ○ were created ○ by dictator ○ Joseph Stalin ○ who ○ assiduously ○ gathered ○ comprimising material ○ on his opposition ○ and colleagues.

Question 2

Revise the sentence below so that the agent is in the subject position of a complete sentence. Use the same verb tense in your revision as in the original. If the agent is unstated, supply a plausible one. Try to retain as much of the original sentence's meaning as you can.

Various directions for future research are suggested by this study.

Answer space

Figure 5.2: The two problems for which participants provided three explanations in the writing experiment, with the one on the right being more difficult

5.2.2 Categorization of Explanations

We collected three explanations from each of the 40 participants in the math experiment, for a total of 120, and three explanations from each of the 30 participants in the writing experiment, for a total of 90. Overall there were 210 explanations, where each explanation is defined as the full text provided by a participant into the answer space. These mostly consisted of sentence fragments or full sentences, but there were several that had multiple sentences. Such explanations were still treated as a single unit, to which the best fitting code was applied [76].

Using data collected from a brief pilot study, two researchers followed the process in [56] to develop a codebook from the explanations in the math experiment, and a separate codebook for the writing experiment. This involved assigning the participant explanations to a set of codes based on their interpreted meaning. These codebooks were iteratively refined until agreement on the codes was achieved. Two research assistants then applied the codebook to the pilot data and discussed discrepancies, seeking clarity for any codes they were unfamiliar with. Table 5.1 shows the finalized version of the codebook applied to the collected math and writing explanation data. The codebook was then applied to the full dataset from each domain by the two research assistants. Next, we measured the code agreement via Inter-Rater Reliability (IRR). The coders achieved a Cohen’s kappa κ = 0.813 for the math experiment and κ = 0.839 for the writing experiment, which indicates a high level of agreement [124].

Code	Definition	Example Explanation
Math experiment		
Calculation	Mentions the computational aspects involved in the problem, e.g., subtraction or use of area	“Because they don't know how to calculate the area”
Clarity-Shape	Relates to the understanding of the depicted shape.	“It may be less clear which part should be calculated because of shading”
Clarity-Text	Relates to the understanding of the text.	“Wording is kinda confusing”
Complexity	Claiming that one problem is more complicated than the other, without further clarification.	“Problem two is more complicated than problem one”
Composite	Addresses an embedded shape used in the problem.	“The picture itself shows other objects such as windows and this might throw off the student.”
Content	General remarks about the problem content that are not captured by other content subcategories.	“The numbers displayed have decimal points”
Meta	A mention of general skills needed to solve any type of word problem, such as focusing, reading, and attention.	“It takes more time to read in problem 2 so students are more prone to getting discouraged”
N/A	Does not provide any sensible explanation.	“340”
Shape-Layout	Mentions the visual element of the word problem's shapes.	“It is more difficult based on the shapes presented in question two”
Step-Num	Indicates one problem requires a certain number of steps / more steps.	“There are more steps to complete in problem 2”
Value-Num	Indicates one problem has more variables/values to work with.	“It has more variables”
Writing experiment		
Answer #	Relating to the number of	“In option one there is only one

	answer choices present in the question.	right answer”
Complexity	Discusses the general difficulty/complexity.	“More complex knowledge needed”
Content	Touches on the content of the question.	“They have to revise it instead of just saying what is wrong”
Meta	Describing a skill required by similar problems, at a more meta level.	“It is hard to write”
N/A	Not applicable or relevant.	“Poor communication with suppliers”
Prework	Discusses the prior knowledge or prework that might be required to answer.	“The second isn't explained in the coursework”
Question-type	Addresses the question's type (MCQ or free response) in the explanation.	“Written answer instead of multiple choice”
Question-text	Mentions the question's text in some capacity, e.g., longer/confusing	“Sentence 2 is more vague”
Rules	Mentions the rules a student would need to know to solve the problem.	“Problem one only requires an understanding of grammar”
Technical	Mentions a specific technical term that might be required to answer.	“In problem two, the subject is not in the beginning of the sentence”

Table 5.1: Coding dictionary for the math and writing experiment responses

5.2.3 Topic Modeling Explanations

Topic models estimate latent topics in a document from word occurrence frequencies, based on the assumption that certain words will appear depending on potential topics in the text. We used two topic modeling techniques, Latent Dirichlet Analysis (LDA [30]) and Non-negative Matrix Factorization (NMF [126]), to further analyze the explanations. LDA maps all documents, in this case the explanations, to a set number of topics in a way such that the words in each document are captured by the topics [9]. NMF uses linear algebra for topic modeling by identifying the latent structure in data, the explanations, represented as a non-negative matrix [142]. The explanation text was lemmatized and stop words were removed, using a common NLP library in Python [29]. No further text processing was performed on the explanation data before running them through the models, as we wanted results without fine-tuning any parameters or heavily processing the data. The results of the topic models were then evaluated against the researcher generated codes, categorizations, and the expert generated KCs for the problems, in order to gauge their effectiveness for this task.

5.3 Results

5.3.1 Crowdworker Explanations

From the coded explanations in the math and writing experiments, we constructed a set of themes, shown in Table 5.2, formed by grouping several of the related codes within each experiment together [201]. In the math experiment the first three themes, Greater Quantity, Shapes Present, and Domain Knowledge, all comprise explanations which address features of the given problems and are indicative of a KC required to solve the problem. Explanations that are grouped into these three themes can be translated into KCs that fit the problem and are indicative of the underlying skill(s) required to solve it. However, the only explanations that suggested a KC that matched any of the expert ones (Compose-by-addition, Subtract, and Rectangle-area) came from the *Calculation* code. The fourth theme, Clarity/Confusion, pertains to the problem's question text or visuals being unclear and hard to decipher. This theme contains explanations that relate to what makes the problems particularly difficult outside of the knowledge required to solve it; from these explanations, one could also derive ways to improve the assessment, such as making the question text more explicit or clarifying the depicted image. The fifth theme, Irrelevant, holds the remaining explanations – those that do not address the problem in a meaningful way, i.e., they are too general or abstract.

Theme (# of explanations)		Codes	KC	Improvement
Math				
Greater Quantity	27	Step-num, Value-num	✔	
Shapes Present	30	Shape-layout, Composite	✔	
Domain Knowledge	33	Content, Calculation	✔	
Clarity/Confusion	15	Clarity-text, Clarity-shape		✔
Irrelevant	15	Complexity, Meta, N/A		
Writing				
Process to Solve	13	Rules, Content	✔	
Domain Knowledge	07	Prework, Technical	✔	
Question Specific Attributes	42	Question-text, Question-type, Answer-num		✔
Irrelevant	28	Complexity, Meta, N/A		

Table 5.2: Themes for the math (above) and writing (below) experiments created from the coded data and if the theme is akin to a KC or an area of problem improvement

In the writing experiment the first two themes, Process to Solve and Domain Knowledge, are indicative of KCs that were required to solve the problems. The only explanations that matched any of the expert generated KCs (Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form) for the problems came from the *Rules* and *Technical* codes. The third theme, Question Specific Attributes, discusses the relative level of difficulty between problems, due to one being multiple-choice and the other being free-response, or the question text differences between the two. This theme relates explanations that address ways to improve the assessment, such as simplifying the answer choices. Finally, the Irrelevant theme again consists of explanations that are not meaningful or overly general.

5.3.2 Topic Modeling to Identify KCs

The 10 topics identified by both the LDA and NMF models, along with the five most common words associated with them, are presented in Table 5.3. From the math experiment data, both the LDA and NMF models had comparable results to one another. They share the same set of topic interpretations and an equally low number of N/A topics. While certain topics in both models are attributed to KC codes, it would be challenging to discern the explicit KC just from the terms. The three primary themes across the ten topics from each model are calculation of area, the visual nature of the shapes in the problems' figures, and how one problem is generally more complicated than the other. We expected some of the expert-generated KCs for the math problems (Compose-by-addition, Subtract, & Rectangle-area) to be identifiable in the topics. Surprisingly *'subtract'* was not a top five term for any topic nor was 'area' a term alongside *'rectangle'* for any topics.

Topic #	LDA Terms	LDA Topic Interpretation	NMF Terms	NMF Topic Interpretation
Math Experiment				
1	figure, question, hard, shape, confusing	Clarity-Shape	problem, longer, figure, steps, lines	Step-Num
2	problem, complicated, 1, complex, 2	Complexity	area, windows, given, figure, door	Calculation
3	step, calculation, need, require,	Step-Num	confusing, wording, question, painted, wall	Complexity

	work			
4	consider, answer, visually, complicated, simple	Shape-Layout	shapes, deal, irregular, question, rectangles	Shape-Layout
5	width, 223, calculate, problem, attention	Calculation	numbers, deal, size, work, need	N/A
6	area, complicated, window, 143, 2	Clarity-Shape	complicated, calculation, somewhat, problem, involves	Complexity
7	confusing, know, abstract, somewhat, term	Complexity	simple, question, involves, consider, shape	Complexity
8	accommodate, time, difficult, shading, shape	Clarity-Shape	harder, visually, figure, shape, make	Clarity-Shape
9	instruction, measurement, equal, forward, straight	N/A	areas, account, figure, need, just	Calculation
10	detail, variable, 340, long, contain	N/A	difficult, calculate, solve, door, width	Calculation
Writing Experiment				
1	answer, prework, specific, pick, confine	Prework	choice, multiple, problem, allows, simple	Question-type
2	multiple, choice, 1, problem, thinking	Question-type	sentence, meaning, needs, subject, problem	Rules
3	sentence,vague, problem, option, right	Question-text	problem, requires, understanding, rules, thinking	Meta
4	long, response, 1, free, variable	Question-type	answer, free, easier, pick, right	Question-type
5	know, comment, paraphrase, range, contain	Rules	people, writing, hard, write, questions	Meta
6	people, write, simplified, question, multiple	N/A	comments, written, eliminate, like, level	N/A

7	need, complex, written, knowledge, number	Complexity	know, subject, verb, tense, agent	Technical
8	comment, problem, choice, multiple, complex	Question-type	answers, correct, just, questions, incorrect	Question-type
9	comment, clause, look, agent, suggest	Technical	clause, concept, agent, ended, like	Technical
10	concept, rewrite, choose, sentence, end	Content	complex, concept, written, ended, like	Complexity

Table 5.3: Top 5 terms from 10 topics identified by the LDA and NMF topic models

Similar to the math topics, both the LDA and NMF models produced comparable results for the writing experiment, with slightly different terms used for the topics between the two. The predominant topic in both models is related to the question type, which is appropriate as it was a dominating category from the qualitative coding. Interestingly, there are not as many topics involving *Complexity* or *N/A*, both irrelevant codes that attribute little to no meaning. The majority of the topics focus on the high-level features of the questions, such as the wording or type. Topic 9 from the LDA model and topic 7 from the NMF one include vocabulary used in two of the expert generated KCs (Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form). However, these topics and the others are not interpretable enough to discern such KCs explicitly from the terms.

5.3.3 Explanation Insights

In addition to some of the explanations being indicative of a KC, such as ones that fall into the *Calculation* or *Technical* codes, many of the other explanations suggested complications with the word problems. In the math experiment, 15 of the 120 total explanations (12.5%) fall into the *Clarity/Confusion* theme from Table 5.2. Additionally, only 15 of the 120 (12.5%) were deemed *Irrelevant* to the problems, meaning that in general the majority of the explanations were either suggestive of an improvement that could be made or a KC required to solve them. The writing experiment had a greater number of explanations, 42 out of 90 (46.67%), that fell into the *Question Specific Attributes* theme in Table 5.2. Only 28 of the 90 (31.11%) explanations in this experiment were deemed Irrelevant to the problems.

5.4 Discussion

Firstly, we wanted to see if the provided explanations could be used to generate fitting KCs for the problems. We found that many of the provided explanations did address the underlying concepts required to solve a problem, more so in the math domain than the writing domain. For example, explanations from the math experiment in the *Greater Quantity* theme often discuss how one problem required the area calculation of more shapes than the other. Solving a problem that involves the area of multiple shapes instead of just a single one has been identified as a knowledge component for similar problems from a previous study [216]. This type of difficulty may be overlooked due to expert blindspot, as the explicit steps taken to solve a problem can get grouped together when it becomes second nature [171]. Eliciting the crowd for explanations such as these can help bring in a diverse level of knowledge, ranging from novice to expert, that can help to make this KC explicit.

From the writing experiment, the *Process* to *Solve* theme consists of the most KC indicative explanations. These often discuss a step required to solve one of the problems, which was usually at the granularity that would make it a fitting KC. Unfortunately the explanations contributed by participants that were indicative of KCs were relatively rare, making up only 20 of 90 (22.22%) of the total explanations from the writing data, compared to 73 of 120 (60.83%) from the math domain. We attribute this difference between domains due to the knowledge required for them, as the math problems were from a middle school class and the writing questions from a college-level writing course.

The two topic models were only able to identify a few topics, each relating to *Calculation*, that fit into a code indicative of a KC that matched one an expert generated. While the terms for the topics can be gleaned for words that suggest a KC such as “area” or “window”, they still lack interpretability and a direct translation into a KC. This is also true of the two models’ results in the writing domain, which identified several topics relating to the *Rule* and *Technical* codes. Without further interpretation, the terms suggest some vocabulary used in the problems, but they are insufficient to derive an actionable KC without further human processing.

Secondly, we wanted to see if the explanations provided insights into how the assessment items might be improved. Both experiments had one theme directly related to improving the surface level features of the problems, such as the question text or images. For instance, in the math experiment, the theme *Clarity/Confusion* addresses the confusion caused by the visual elements of the problems. The included images for the questions are a key aspect to the assessment and beneficial to problem solving, but may be misinterpreted in a way the content creators may not have intended [71]. Correcting the images can allow for better assessments; based on the explanations we received, a student may answer incorrectly purely based on the poor image design. Across both domains, the 10 topics identified by each model are mostly those that indicate areas of problem improvement. While the models performed poorly at generating KCs from the explanations, many of the topics and terms were indicative of student struggle due to confusion with the text or image of the problems. In total, 12.5% of the explanations in math and 31.11% in writing were considered irrelevant to the task and presented problems. Even with limited instruction and the varying backgrounds, participants were able to provide insights into the problems that could be used for baseline KC generation or identifying areas of assessment refinement.

5.5 Future Work

For future work, we plan to integrate this process in a learner-sourced context, where participants (i.e., students) potentially have more commitment and domain knowledge that could be leveraged [183]. This would enable us to properly train them to provide such explanations throughout the course, rather than completing the task once with only a brief instruction like the crowdworkers did in this study. Ultimately, we envision a workflow in which students submit explanations for why certain problems are difficult; these explanations are then peer reviewed and presented to the teachers (or relevant parties) to help them identify potential KCs and improve the assessment items. This procedure is analogous to the find-fix-verify pattern in crowdsourcing, which has been shown to be effective [21]. However, before reaching this point, the interpretability of the models will need to be improved or another technique should be utilized. This study demonstrates a first step in developing such a workflow, providing initial insights into how crowdsourced explanations might be leveraged for KC generation and assessment content refinement.

5.6 Conclusion

In this study, we gathered explanations for the relative difficulty between two mathematics questions and between two English writing questions from crowdworkers. We found that crowdworkers were able to generate valuable explanations that were indicative of a KC required to solve the problems or a suggestion for how to make the problems clearer. Understandably, they were able to provide better explanations in the easier domain of middle school math than in an undergraduate English writing domain. However, in both experiments, a majority of the explanations either pertained to identifying a KC or area of improvement, rather than being irrelevant. The LDA and NMF models created topics akin to the researcher generated codes, although the interpretability of these topics based solely on the terms is limited in usefulness. Nevertheless, the categories from the coding and topic models ultimately assisted in clustering explanations that were either indicative of a KC or an aspect of the problem that could be improved.

Chapter 6
Learnersourcing Skill Tags for Assessments

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Huy A. Nguyen, and John Stamper. "Leveraging students to generate skill tags that inform learning analytics." In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 791-798*. International Society of the Learning Sciences, 2022.

6.1 Introduction

Educational technologies, such as intelligent tutoring systems and digital learning platforms, often employ a mapping of skills to assessment items in order to measure student performance and guide them in the learning process [26]. These skills treat student knowledge as “an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks” [116]. Typically the skills mapped to assessments in these technologies are at a fine granularity intended to measure the low level concepts assessed by a given activity, compared to the higher level skills often found in summative assessments that are often at the standard or learning objective level [84]. This skill mapping is intended to capture the student’s underlying cognitive process and is vital to many core functionalities of educational software, enabling features such as adaptive feedback and hints [167].

While having these skills allows for student modeling and can assist in analytical pipelines like open learner models or learning dashboards, developing the mapping of skills to assessments poses several challenges. Traditionally an expert is employed to create the mapping of skills to an assessment, potentially with assistance from an instructional designer or learning engineer [115]. This process can be arduous and often requires methods such as think-alouds to elicit the skills from the expert while avoiding pitfalls such as expert blindspot [171]. As a result, scaling this process is challenging due to the time constraints, resources for eliciting the skills, and recruiting the domain experts to assist with skill tagging [145].

Attempts have been made to scale this process, as several studies have turned to crowdsourcing to recruit a pool of humans to skill tag assessments [157, 161]. While there was surface level success with this pool of human judgment, prior knowledge of the domain is often required to properly generate and map skills to assessments [115]. Enlisting students, who have varying domain knowledge and have been successful in related tasks, may offer a viable solution to this problem [233].

To explore how we could potentially engage and scale a group of humans in the skill tagging process, we sought to test a novel solution that utilized a more knowledgeable base. To utilize students in generating skill tags, we deployed a completely optional activity where students could input three skills, via free-text boxes, needed to solve a problem in the context of four undergraduate online courses. This study contributes new knowledge about student participation with optional tasks in an online course, particularly one involving skill tagging. It also supports how student performance correlates with the quality of their skill tags. The results provide insights into the feasibility of leveraging students to skill tag problems and challenge us to further investigate how we can deploy interventions that leverage student knowledge to develop stronger learning analytics.

6.2 Methods

The present study takes place in a digital courseware platform known as the Open Learning Initiative (OLI). OLI is an open-ended learning environment that offers courses from a variety of domains (such as chemistry, biology, statistics, economics, etc.) and consists of interactive activities and diverse multimedia content [27]. OLI activities are presented in two distinct categories: low-stakes/formative, providing students with feedback, or high-stakes/summative, used to evaluate student learning at the end of a structured unit. The low-stakes formative assessments in the system are all optional, allowing the students to scroll by them and focus purely on the instructional content. These assessments consist of a variety of question types such as multiple-choice questions, short answer, and dropdown style questions. Each question in OLI is broken down into one or more problem steps, where each step corresponds to an opportunity for student input. For instance, if a question asks a student to set the value of three dropdown boxes, then it consists of three unique steps that each have their own set of feedback and correct or incorrect responses. This distinction between a problem and its steps is important, since students may work on a problem, but not complete all of the steps due to reasons such as not knowing the answer to some of them, wanting to skip them to save time, or getting distracted [42].

6.2.1 Context and students

For this study, we used data collected from four instances of two different introductory courses, one in chemistry and the other in Python programming. The two instances of the introductory chemistry course were taught at a community college in the western United States. This course provides students with fundamental knowledge of chemistry concepts, preparing them for future biology and chemistry courses. The course is generally geared towards freshman and sophomore undergraduates from varying degree backgrounds, with a majority of the students pursuing a chemistry-related degree. Our data comes from the fall semester of 2020, when the introductory chemistry course was offered in the OLI system.

The OLI content the students used for these two instances of the chemistry course in this study covers the topic of elements and compounds and consists of thirteen separate modules. Each module consists of several topic headers, containing paragraphs of instructional text and low-stakes activities embedded throughout. There are a total of 33 low-stakes and completely optional activities embedded throughout the thirteen modules of the course, not including the skill tagging activity used for this study. These activities include multiple-choice questions, selecting the correct option from a dropdown, drag-and-drop exercises, and submitting a short answer to compare against an expert response. Each of these activities is broken down into steps, depending on the components of the activity, for a total of 178 unique steps. For instance, if a problem has three fill-in-the-blank boxes, then that problem would consist of three unique steps. Every activity and their steps provide students with feedback after they have been answered. Additionally, students have unlimited attempts to answer these questions, so they can continue until they are correct or choose to advance, regardless of a correct or incorrect response.

The two instances of the introductory Python programming course were taught at an R1 university in the northeastern United States. This course provides students with knowledge on introductory concepts in the Python programming language, preparing them for their future coursework in computer science. This course was taught to incoming master's students in a human-computer interaction program who indicated by self-report they did have a sufficient understanding of programming concepts. The collected data we used comes from the summer semesters of 2020 and 2021, when the course was offered in the OLI system. The OLI content in the two instances of the programming course covers the topic of iteration and consists of twelve separate modules. There are a total of 41 low-stakes and completely optional activities embedded throughout the twelve modules of the course. Each of these activities is broken down into steps, depending on the activity, for a total of 113 unique steps.

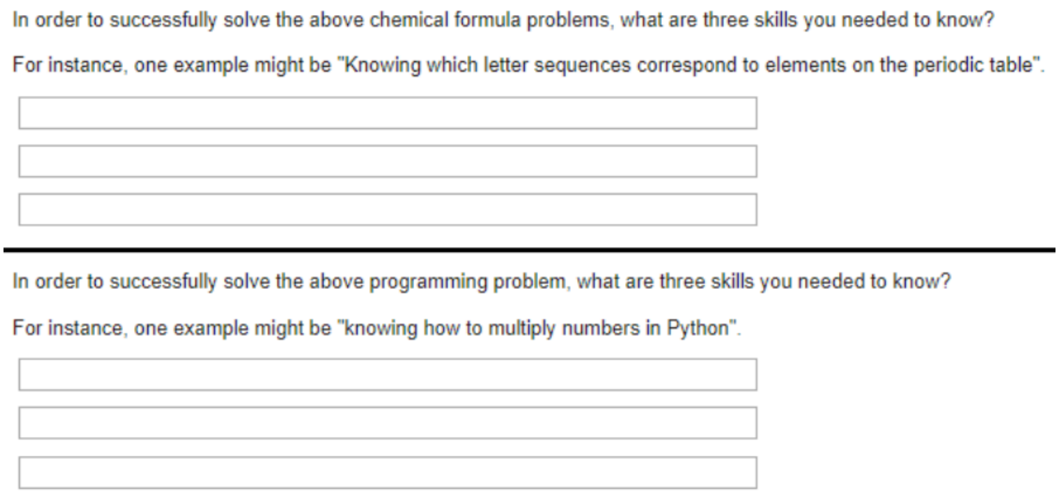
While the four course instances were taught in different semesters by different instructors, students were provided with the same set of instructions regarding the use of the OLI materials. They were not required to answer the questions found throughout the OLI modules or even access them. Students were provided with an "Introduction to OLI" module, which is an overview of how to effectively make use of the system and the concepts that will be covered in the course. All the instructional materials in OLI were optional to the students; there was no requirement for them to access or complete the materials. However, students were assessed on the concepts covered by the OLI materials, so it was beneficial for the students to utilize them. A further breakdown of the course offerings, including the anonymized instructor, semester, and number of students that accessed the course materials can be found in Table 6.1.

Course	Semester	Instructor	Student Count	Number of Activities
chem 1a	fall 2020	t1	23	33
chem 1b	fall 2020	t2	26	33
programming 1a	summer 2020	t3	33	41
programming 1b	summer 2021	t4	30	41

Table 6.1: The four course instances used in this study

6.2.2 Data collection

We focus on an activity we added to this course that involves the students generating three skills that they believe are required to solve a particular problem in the course1 . In the chemistry course, this activity is found in the ninth module of the OLI content for this section of the course. This module provides several paragraphs of instructional text, two worked examples, and several low-stakes multiple-choice questions on the topic of nomenclature for ionic compounds with polyatomic ions, along with this single activity. In the programming course, this activity is found in the fourth module of the OLI course. This module covers the concept of python for-loops, consisting of several paragraphs of instructional text, multiple-choice questions, and two small programming exercises. In both courses, the activity is presented in the same low-stakes and optional format as the other activities found prior in the course. This task that prompts students to generate the three skills can be found in Figure 6.1. In the activity, students are prompted to generate three skills that are needed to solve the problem stated above this task. The students input the text in three different free-text boxes. The instructions for the task are minimal to encourage student participation, as lengthier text might deter students [204]. We also provide them with a domain contextual example of what a skill might look like for a different, non-related, problem in1 the course. Aside from that, no training or scaffolding was provided to the students to help them generate three skills for the question. We intentionally wanted to keep this low-stakes and optional, to examine the students' participation with the task and the quality of their contribution.



In order to successfully solve the above chemical formula problems, what are three skills you needed to know?

For instance, one example might be "Knowing which letter sequences correspond to elements on the periodic table".

In order to successfully solve the above programming problem, what are three skills you needed to know?

For instance, one example might be "knowing how to multiply numbers in Python".

Figure 6.1: The self-explanation activity and accompanying question presented to the students

Student data was collected from their interactions with the 33 activities found in the chemistry course and the 41 activities in the programming course, in addition to the skill generation task. Since the skill generation task is our outcome, we focus our analysis on the other activities that the students completed in the courses, which consisted of a total of 178 unique steps in the chemistry courses and 113 unique steps in the programming ones. On average, an activity in the course consists of 3 unique steps, such as a single activity having the student select from three different dropdown menus. All of the activities found in the OLI course were completely optional; students could do as much or as little as they desired. For instance, sometimes a student would begin working on an activity, but did not complete all of the parts in it. As a result, the system logs them having worked on that activity and also provides the exact number of steps that they completed. For this data set in particular, it is more common for students to fully complete an activity if they start it, i.e., they will attempt all of the steps of a problem.

Our data also consists of three metrics related to student performance on the activities. When a student works on a step for a given activity, OLI records if their first attempt at that step was correct or not. A first attempt at a problem can be a strong indicator of a student's current understanding of the concepts being assessed [48]. Relatedly, the total number of incorrect attempts made at a given step and the total number of correct attempts are recorded. These numbers can potentially exceed the total step count, as a student could correctly answer a question, then select an incorrect answer to see the feedback, then select the correct response once again, registering two correct and one incorrect attempt for that step.

6.2.3 Analysis

In order to assess the quality of the three student-generated skills, we had two item-raters evaluate each skill to determine how much it fit the problem, if it was at the appropriate granularity, and its match to the three skills previously assigned to it by other domain experts. Both raters had content-area expertise, ample experience in skill tagging, and experience applying coding schemes to student data. After familiarizing themselves with the OLI course module once again, they were instructed to go through each skill one at a time and place it into one of four categories. The inter-rater reliability was calculated, and the Cronbach's alpha value was .94, as the raters only disagreed on the categorization of a few skills. These discordant skills were discussed among the two raters until they reached a consensus on the categorization of them using the coding categories. Each of these categories was assigned to a numerical ranking (1-4) that also represented the goodness of the student-contributed skill. In this instance, a ranking of 1 was the best and indicated that the student-generated skill matched the expert-generated skill. A ranking of 4 was the worst and indicated that the student-generated skill was not relevant to the problem and therefore it did not match. A full description of the four categories can be seen in Table 6.2. These four categorizations are based evaluation rubrics from previous studies for assessing math problem solving skills and evaluating knowledge concept maps from novices and experts [73, 111].

Category	Rank	Description	Examples
Expert Match	1	The skill matches one of the three skills originally tagged to the problem by an expert.	"Combine ions in the smallest ratio"
Match, Not Granular	2	The skill is very similar to one of the three expert-generated ones, however it could be more specific.	"Know charges on polyatomic ions"
Problem Relevant	3	The skill is technically utilized in the problem, but it is not necessarily what is being assessed given the context.	"Write the formula that reflects this ratio"
No Match	4	The skill is not relevant to the problem, it is not being utilized or assessed by it at all.	"Knowing nomenclature"

Table 6.2: The four categories the student-generated skills were placed into along with an example student generated skill from the chemistry course

The two problems used in each of the courses were selected due to previous student data collected during their use indicating that students tend to struggle on both of them, with 40% of students getting an incorrect answer on their first attempt. Additionally, they each had three skills previously assigned to them by domain experts upon the creation of the courses. Our two item-raters verified that these three skills were appropriate for the problems and at the correct granularity for what was being assessed. While having experts evaluate student-generated skills is not scalable, in the scope of this study, we want to first investigate the feasibility of having students generate such skills. To assess the quality of these student-generated skills, we need to compare their outputs to that of experts that typically do this task. A list of the three skills tagged to each problem can be found in Table 6.3.

Expert Skills - Chemistry	Expert Skills - Programming
1. Identify the cation and anion and their charges	1. Iteration over a value using the range() function
2. Write chemical formulas for ionic compounds that contain polyatomic ions	2. Utilizing a print statement in conjunction with numerical statements
3. Combine the ions in the smallest whole number ratio	3. Modify an existing program to fit a new set of instructions

Table 6.3: Three expert-generated skills for chemistry (left) and programming (right)

6.3 Results

The student-generated skills were evaluated by experts to determine their quality. Then we analyzed how the student interactions in the course correlated with both student participation on the task and the quality of their contribution. We leveraged measures of central tendency to report the varying categories of the student-generated skills. Then we investigated the different patterns of student participation and performance in the course by looking at their interactions with the varying low-stakes activities and their steps embedded throughout the course. A Bonferroni correction was applied to post hoc analyses that follow [15].

6.3.1 Students Generating Accurate Skill Tags

Across all four course instances used in this study, a total of 112 students accessed the OLI course. Among those 112 students, 64 of them completed the optional skill generation task where they generated three skills for the presented problem. To assess the quality of these student-generated skills, we had two expert item-raters evaluate all 192 of their contributions. The raters checked if the student-generated skills matched an expert-generated one, indicating that they were accurate, meaning they are written at the correct level of specificity and model part of the knowledge required to solve the problem. This evaluation revealed that 23 (11.98%) of the contributed skills matched an expert-provided item and 65 (33.85%) matched a skill, but needed to be refined for further granularity. Table 6.4 shows the further breakdown of the item-rater categorizations of the 192 student-generated skills into the four different categories. Over half (54.17%) of the student-generated skills fell into the bottom two categories, which indicates that even with revision they are not usable or particularly useful. All of the skills labeled as the fourth ranking, “No Match”, were one word or nonsensical responses by the students for the activity.

Course	Student Skill Count	Expert Match	Match, Not Granular	Problem Relevant	No Match
chem 1a	36	7	15	6	9
chem 1b	54	13	24	14	3
programming 1a	54	1	14	29	10
programming 1b	48	2	12	24	10
Totals	192	23	65	72	32
(%)		(11.98%)	(33.85%)	(37.50%)	(16.67%)

Table 6.4: The count of the four ranking categories assigned to skills in each of the courses.

Each of the activities in the two course domains had three skills tagged to it by domain experts, previously shown in Table 6.3. Student-generated skills from the chemistry courses were able to match all three of the expert skills for the problem, while only two of three expert skills were matched by students in the programming course. Table 6.5 shows an example of a student-generated skill from both domains that was categorized as an expert match by the two item-raters.

Expert Skill - Chemistry	Expert Skill - Programming
Identifying the cation and anion and their charges	Knowing how to use a print statement in conjunction with numerical statements
Student Skill - Chemistry	**Student Skill - Programming**
Identifying the cation, anion, and the charges	Knowing how to manipulate equation to have odd numbers printed instead of even numbers

Table 6.5: An example of a student-generated skill that matches one of the chemistry expert skills (left) and one of the programming expert skills (right)

6.3.2 Student Participation and Performance

In total, of the 112 students that were part of four instances of the courses, 19 of them did not access any of the course materials in the given modules. A total of 49 students were enrolled in the two chemistry courses, where 38 students accessed the materials, and 30 students did the skill generation activity. The chemistry course consists of a total of 33 optional low-stakes activities, not including the skill generation one, and on average the students completed 26.38 of the 33 (79.93%) activities. The two programming courses had 63 students enrolled in the courses, where 59 students accessed the materials, and 34 students did the generation activity. The programming course consists of 41 optional activities, not including the skill generation one, and on average students completed 30.48 of the 41 (74.34%) activities.

To determine which features of student interaction in the course were indicative of their participation in the skill generation activity, we performed a series of unpaired t-tests on their behavior with the other low-stakes activities found in the courses. This revealed a significant difference between the student participation with the other activities in both the chemistry and programming course and their participation in the skill generation task. An unpaired t-test showed there was a significant difference in the number of activity steps completed by students in the chemistry course who did the skill generation tasks ($M = 164.57$, $SD = 20.11$) and those that did not do the task ($M = 39.47$, $SD = 52.45$), $t(47) = 11.819$, $p < .0001$. A similar result was found for the programming course, with students doing the skill generation task ($M = 111.03$, $SD = 18.41$) and those that did not do the task ($M = 54.00$, $SD = 40.27$), $t(61) = 7.699$, $p < .0001$. Students who did the skill generation task were also more likely to complete all of the steps present in the activities embedded throughout the course. Similar significant results were observed for the number of activities done by a student and their participation for the skill generation task in the chemistry course, $t(47) = 11.483$, $p < .0001$ and the programming course $t(61) = 6.299$, $p < .0001$. This result supports the previous one, as the activities found throughout the course are composed of multiple steps and a subset of students completed all the 33 or 41 low-stakes activities respectively.

We then investigated whether student interactions with the other low-stakes activities in the course correlated with the quality of their generated skill contribution, to see how we might predict or promote better skills from the students. While there was a significant difference found between student participation in the skill generation task and participation in the other low-stakes activities throughout the course, it was not found to significantly correlate with the quality of the student contribution in the chemistry course ($\rho = 0.18$, $p = 0.333$) or programming course ($\rho = 0.18$, $p = 0.303$). Additionally, student performance on the activities as measured by their first attempt correct on the problems was not significantly correlated with the quality, measured by itemraters categorization, of their skill contributions for the chemistry course ($\rho = 0.29$, $p = 0.119$) or programming course ($\rho = 0.09$, $p = 0.615$).

6.4 Discussion

In this research, we investigated how we might enlist students working through online courses to assist in the skill tagging process. We found that even with the task being optional and only providing brief instructions with no scaffolding, some students were able to generate skills that matched experts, ones that could be utilized without any modifications. However, a majority of the generated skills were not sufficiently detailed, and even with revision, they would not be suited for use. The students who chose to participate in the task typically completed all of the other optional activities found in the course. In exploring what features of student interaction in the course correlated the quality of the skills they generated, we did not find any significant correlations with their performance on the other low-stakes activities. These findings suggest that students can generate and tag expert level skills to problems from an optional and low-stakes activity within an online learning environment, but they might need more scaffolding to consistently do it.

Evaluation of the 192 student-generated skills indicated that, while not the majority, students across both course domains generated skills that matched all three expert ones in chemistry and two out of the three for programming. Students in all four course instances generated many skills that were categorized as a match but needed a bit more granularity to directly match an expert-generated skill. While these levels of skills cannot be as readily used as the direct match ones, they could still serve as an initial baseline to feed into a learning analytics system. We are hopeful that through more instruction or guidance, we can increase the quality of the student contributions by having them think deeply about the granular details of the skills they use to answer the activities.

With all the low-stakes activities embedded throughout the course being completely optional, including the skill generation one, there was still a high amount of overall participation from the students. This was particularly surprising for an optional activity, which generally has lower participation rates due to the lesser perceived value students see in completion of the activity [77]. While past skill generation methods have relied on experts (e.g., [115], our study presented the task as a lowstakes activity, seemingly fitting in among the fill-in-the-blank and drag-and-drop activities found on the other modules of course content. Leveraging just the native features of the system, in this case free-text boxes for short answer questions, we were able to provide students with the skill generation task seamlessly and without requiring them to utilize yet another platform. It is likely participation would be even greater if the task was required by students or embedded into a high-stakes assessment, such as a quiz question. However, this approach would introduce another series of potential complications, such as requiring it to be graded and potentially introducing an abundance of unacceptable questions contributed by students that do not wish to do the activity but are forced to in the context.

6.5 Limitations

Our contributions should be interpreted against the following limitations. First, our study was conducted across two domains, chemistry, and programming. As skill tagging is directly influenced by the domain of the problems, this may impact the generalizability of the results. Second, students from each domain skill tagged a single problem in each instance of the course. Similar to the domain, the content of the problem is directly related to skills it assesses, so our results might differ depending on the problem or domain. Finally, we did not heavily investigate the background of the students doing the task. It is possible that some students were more familiar with the concepts of skill tagging or articulating this process more than others.

6.6 Conclusion

This research demonstrates a first attempt at directly engaging students in the processes of skill tagging problems across the domains of chemistry and programming. Our results highlight how student behaviors regarding their completion of activities in the course are indicative of their participation on the optional skill tagging task. Requiring the student to do the activity, such as putting it in a high-stakes assessment or having students use another system, is not necessarily required for contributions, some of which are evaluated as being on par with expert ones. Although rare, students can provide expert-level skills, without detailed instructions, prior training, or scaffolding. However, most of the skills students generated would not be usable, even with revision. As a result, we do not recommend enlisting students in the skill tagging process through this mechanism, as they might require training or better scaffolding to submit usable skills. This research demonstrates an attempt to help scale online learning analytics and improve educational resources, by leveraging the students in a course. This work opens further opportunities for both engaging students in the process of skill tagging and promoting their behavior that leads to a higher quality contribution for future tasks. Future research should consider expanding domains and experimenting with different amounts of training or scaffolding for the task.

Chapter 7
Equitable Participation in Learnersourcing

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. "Who writes tomorrow's learning activities? Exploring community college student participation in learnersourcing." In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023, pp. 664-671*. International Society of the Learning Sciences, 2023.

7.1 Introduction

Having students develop assessment questions has a long history as a learning activity, one that has shown real benefit in supporting student learning [4]. These types of activities integrate deep engagement around subject matter with critical thinking and creative practices [57]. Through the instrumentation of this process, student engagement can be leveraged in ways that provide meaningful data around student interaction, in addition to new student-generated learning assets that can support future learners [61]. This is known as a form of learnersourcing, where students complete activities that produce content that can be leveraged by future learners [110]. The continual creation and improvement of these questions allows for a greater breadth of topic coverage, helps to identify well-constructed and valid assessments, and as a result, enables improved learning opportunities.

While asking students to write new quiz or exam questions is a time-honored approach in many classrooms, current learnersourcing investigations emphasize the online context, where students' efforts to master domain content within a digital learning environment can be effectively studied at scale [81]. Specific implementation of learnersourcing activities can vary greatly between instructors, however, particularly in whether completion of these activities are treated as mandatory or optional [162]. This distinction between mandatory and voluntary implementation is important: students who are offered a choice in completing learnersourcing tasks perceive these activities as having a greater value, gain more autonomy in the course, and contribute higher-quality questions compared to students that are required to participate [213]. Indeed, efforts to force student engagement may backfire; requiring these tasks can lead to student disengagement as they participate with minimal effort in order to satisfy the requirements of the activity [107].

On the other hand, making the activities optional comes with its own risks: the activities may be neglected by the students who could benefit the most from these interventions, as oftentimes only the most driven students choose to participate in optional tasks [96]. This type of self-selection would have an impact beyond the individual students for learnersourcing activities, as only the top performing students may be generating data and new questions. This would in turn influence the question banks, hints, analytics, etc. generated by the students, limiting the diversity of the contributions, creating potential bias in the generated content, and potentially excluding a novice point of view that could be beneficial to learners [171]. Ideally students of all backgrounds and knowledge levels would participate in learnersourcing activities, but previous work has indicated otherwise – that participation for these question generation activities can be as low as less than 5% [59]. These findings are further complicated by where such investigations take place, with a majority of the learnersourcing activities being deployed at top R1 four-year universities around the world [227].

Therefore, more research is needed to investigate which students are participating in these learnersourcing activities, how these interventions work and who they are targeting. To this end, we deployed several optional MCQ generation activities in three online courses across two community colleges in the United States. Accompanying these tasks is a demographic survey to help better understand the students in the courses. As students worked throughout the first four or five weeks of their online course, they were presented with several opportunities to generate a MCQ for the given unit they were working through. We analyzed which students were participating and how their demographics and performance in the course may have influenced their participation. Our work makes the following contributions towards online learning and learnersourcing tasks. First, we provide insights into the attributes of students that participate in learnersourcing activities. Second, we derive a set of performance measures commonly found in online courses that can serve as predictors of student participation in related activities

7.2 Methods

This study was conducted in three different courses at two 2-year community colleges located on the west coast of the United States. All three courses took place online during the fall 2021 semester and IRB approval was received for the survey and activities added to the courses. The three courses were introductory chemistry, advanced chemistry, and introductory statistics. The two chemistry courses were taught at the same community college, but by different instructors. Students taking introductory chemistry were required to have previously passed a course covering the topics of linear algebra. For the advanced chemistry course, students were required to have passed both a linear algebra course and an introductory chemistry course at the college-level. The statistics course was taught at a separate community college by a third instructor. The only prerequisite for students in that course was to have passed a college-level intermediate algebra course.

We utilize data that came from four to five week-long units that were used towards the beginning of each course. This data consists of student interactions in the course along with their performance on the quizzes found at the end of each unit. There were a total of 64 students across all three courses, who were taking the courses to receive credit towards their respective degrees. There were no students enrolled in both of the two chemistry courses. Table 7.1 shows the number of learners in each course, along with a breakdown of their self-reported gender, ethnicity, and first-gen status. It also includes the number of units, and therefore quizzes, in the respective course, as introductory chemistry had five units and advanced chemistry and introductory statistics had four units. Our demographic questions accepted free text input to allow students the highest flexibility in identifying their background.

Course	Units	Students	Male	Female	First-gen	Hispanic/ Latin	Asian	White
Introductory Chemistry	5	17	5	12	8	12	3	2
Advanced Chemistry	4	18	4	14	12	12	5	1
Introductory Statistics	4	29	6	23	19	15	6	8

Table 7.1: Breakdown of the students in each course, their demographic information, and the number of units

All three courses were deployed on the same learning platform, known as the Open Learning Initiative (OLI), which has been used in previous studies involving online learning at community colleges [20, 198]. It contains functionalities akin to popular learning environments often utilized at universities or in MOOCs. Each unit in these courses was equivalent to a chapter in a textbook, consisting of five to ten related topics and taking up roughly one week to cover. The units contain multiple pages of instructional content featuring text and brief instructional videos. These webpages also host multiple low-stakes activities interspersed amongst the instructional content for students to use as practice opportunities. They include multiple choice, short answer, essay, matching, and fill-in-the-blank style questions. All of these activities act as formative assessments, intended to provide students with instructional feedback. As such, they are completely optional and do not account for the students' grade in the course. Additionally, students may make any number of attempts on these activities, receiving instant feedback on their response with each attempt. In all three of the courses, each unit concludes with a summative assessment in the form of a quiz that tests students on the material covered in that unit. The quizzes consisted of only multiple-choice or fill-in-the-blank questions and ranged from 4 to 22 questions. Students' scores across all of the quizzes counted towards a low percentage (5-15%) of their final grade in the course. All student data collected from OLI is securely stored in accordance with its IRB approval. In addition to the OLI platform, students in these courses utilized a learning management system for the other parts of their course, such as submitting homework assignments or viewing announcement posts.

7.2.1 Data Collection and Analysis

Our dataset came from the four to five week-long units at the beginning of each course, with four primary components: 1) Demographic survey, 2) Formative assessments, 3) Summative assessments, 4) Learnersourcing activities.

Demographic survey

When students first accessed the learning environment which hosts the formative and summative assessments, they were prompted with a brief demographic survey to complete. The survey asked the students to specify their gender identity, ethnicity, and if they were a first generational (first-gen) college student in their family. Students that did not fully complete this survey were not included in the present study. Additionally, as many of the responses were free-form, we had two researchers standardize the student responses (e.g., fixing typos), during this process there were no discordant cases.

Formative assessments

Throughout each course there are multiple formative assessments, commonly referred to as problems, embedded amongst the instructional text and videos intended to provide the students with practice opportunities and immediate feedback. They consist of multiple-choice, short answer, essay, matching, and fill-in-the-blank style questions. These activities are optional and do not impact the student's grade in the course. Table 7.2 shows the total number of formative and summative assessments in each course – note that these do not include the count of the MCQ generation activities, which we describe below.

Summative assessments

The end of each unit concludes with a page summarizing the content that was covered in the unit. This page also contains a link to the unit's quiz that students complete for a small percentage of their final grade. It consists only of multiple-choice and fill-in-the-blank style questions that can be automatically graded. In this study, the smallest quiz contains 4 questions, and the largest quiz contains 22 questions

Learnersourcing activities

At the end of each unit in each course, we placed a learnersourcing activity that prompts students to generate an MCQ targeting any concept they learned from the unit. The interface of the MCQ generation activity includes the brief instructions for the students. The two bullet points shown in the activity's instructions reflect the unit's learning objectives which the MCQ should target. The number of MCQ generation activities is equal to the number of units in the course.

Course	Formative Assessments	Summative Assessments
Introductory Chemistry	126	5
Advanced Chemistry	94	4
Introductory Statistics	37	4

Table 7.2: The number of formative and summative (quizzes) assessments in each course

Our primary variable of interest is student participation with the learnersourcing activities in their respective course. In this study, we consider a student as having participated in the learnersourcing activity if they submitted a contribution that contains a question pertaining to the course's learning objectives, a correct answer choice, and three distractor options. If a student submitted a blank response, a random string of characters, or made no submission, they were not counted as having participated in the learnersourcing activity. Note that it was rare for students to exhibit this behavior, as the vast majority of them either skipped the learnersourcing activities or made an honest effort in their contribution to generate a MCQ. To measure student performance on the formative assessments, we used their accuracy on the first attempt they made on the problem. If they correctly answered the problem on their first attempt, then they would have the first-attempt correct for that problem. Previous research indicates that a student's first attempt at a problem is a strong indicator of their knowledge of the material [48]. In the forthcoming analysis we utilize the average quiz scores of the students, as it represents their performance in the course up to that current point in the course.

7.3 Results

To understand which students were participating in the optional learnersourcing tasks, we first analyzed their demographic information in relation to their potential contributions to the learnersourcing activities. Next, we investigated the different patterns of student participation and performance by looking at their interactions with the formative and summative assessments embedded throughout the courses.

7.3.1 Student Demographics

In total, 37 of the 64 (57.81%) students participated in at least one of the learnersourcing activities in their respective courses. To further investigate student participation with the learnersourcing activities in the courses, we looked at the demographics for students that contributed to any of the MCQ generation tasks. A Fisher's exact test revealed that there was no statistically significant association between gender and participation with any of the learnersourcing activities (p=.484). Similarly, there was no significant association between first-gen status and student participation with the learnersourcing activities (p=.794). We also looked at participation on the tasks related to the students' self-reported ethnicity. A chi-square test of independence showed that there was no significant association between ethnicity and task participation, X 2 (2, N=64)=.27, p=.873. Table 7.3 provides the count of students who participated in the learnersourcing activities in each demographic group.

Participated in learnersourcing	Students	Male	Female	First-gen	Hispanic /Latin	Asian	White
Yes	37	9	28	26	23	7	7
No	27	6	21	18	17	6	4

Table 7.3: Student participation with any of the learnersourcing tasks and their demographic information

For the 37 students that participated in at least one or more of the learnersourcing activities, we investigated if their demographic background had any statistically significant effect on the percentage of learnersourcing activities they completed. Note there were four learnersourcing opportunities in advanced chemistry and introductory statistics, and five opportunities in introductory chemistry. An unpaired two tailed t-test revealed that there was no significant effect of gender on the number of learnersourcing activities students worked on, t(35)=.95, p=.348, with females (M=.59, SD=.09) doing slightly fewer of the learnersourcing activities than males (M=69, SD=.10) on average. There was likewise no significant difference in the percentage of learnersourcing activities between first-gen students (M=.62, SD=.08) and others (M=.61, SD=.13), t(35)=-.12, p=.452. A Kruskal-Wallis test was conducted to examine the differences of the students' self-reported ethnicity and the percentage of learnersourcing activities they completed. There was once again no significant differences in participation found between groups, H(2)=0.516, p=.773.

While students' demographic background had no significant association with their participation with the learnersourcing tasks or the amount of learnersourcing tasks they engaged with, we also looked at how this information might be associated with their overall performance and participation with the other material found throughout the course. We found no significant effect of gender on the percentage of other formative assessments done in the course, t(62)=1.48, p=.144, where males (M=.54, SD=.11) and females (M=.41, SD=.09) had similar participation levels. There was likewise no significant effect of gender on the average quiz scores, t(62)=.61, p=.546, with males (M=.73, SD=.05) and females (M=.68, SD=.09) receiving similar scores. Similar null effects were found for the formative assessments, t(62)=-1.07, p=.287, and quiz scores, t(62)=-.83, p=.407, between first-generation students ($M_{formative}$=.46, $SD_{formative}$=.09; M_{quiz}=.71, SD_{quiz}=.07) and others ($M_{formative}$=.38, $SD_{formative}$=.11; M_{quiz}=.65, SD_{quiz}=.11). Finally, a Kruskal-Wallis test revealed no significant formative assessment participation, H(2)=3.913, p=.141, or quiz scores, H(2)=1.233, p=.539, between students' self-reported ethnicities.

7.3.2 Student Performance

We focus on how student participation and performance within their respective course might reflect their contribution to the learnersourcing activities. Our study showed that students who participated in the learnersourcing activities (M=.62, SD=.07) had a significantly greater percentage of the formative assessments completed in their respective course than those that did not (M=.18, SD=.02), t(62)=-8.07, p<.005. Relatedly, there was a significant positive correlation between the percentage of formative assessments done by the students with the number of learnersourcing activities they completed, r(62)=.28, p<.005. Table 7.4 provides the average amount of formative assessments completed in each course by students who participated or did not participate in the learnersourcing activities, including those students that only did the quizzes in these averages.

Course	Learnersourcing Participation			
	Average Percentage of Formative Assessments Completed		Average Quiz Scores (out of 100)	
	Yes	No	Yes	No
Introductory Chemistry	73.71	26.98	72.83	67.00
Advanced Chemistry	61.70	10.99	67.50	73.03
Introductory Statistics	49.83	22.45	64.23	71.27

Table 7.4: Average percentage of formative assessments completed in the courses and the average quiz scores, out of 100, by students that participated in the learnersourcing activity (Yes) and those that did not (No)

While participation in the course was positively correlated with doing the learnersourcing activities, as expected, we wanted to further investigate if these activities were more likely to be done by students already performing highly in the course or if it was a true mix of the students. We found that students who performed better on the formative assessments in the course were also more likely to contribute to the learnersourcing activities. These students who participated in the learnersourcing activities (M=.48, SD=.10) compared to those who did not (M=.66, SD=.03) had a higher percentage of correctness on their first attempt in the formative assessments, t(62)=2.99, p<.005. For the 37 students that participated in one of the learnersourcing tasks, there was also a positive correlation between the number of learnersourcing activities completed and the percentage of correctness of first attempts on the formative assessments, r(35)=.35, p<.005.

Next, we examined student performance on the summative quizzes at the end of each unit. Table 7.4 also shows the average quiz scores across all three courses divided into two groups based on if the students participated in any of the learnersourcing activities. We analyzed how a student's performance on the quizzes correlated with the amount of learnersourcing activities they completed. Ultimately, we found a significant positive correlation between a student's average quiz score and the number of learnersourcing activities they did in the course, r(s)=.26, p<.05. Interestingly, across all three courses, seven students had a perfect quiz average, receiving full credit for all four or five quizzes depending on the course. However, of those seven students, only one of them participated in the learnersourcing activities, contributing to all four of them in the advanced chemistry course.

7.4 Discussion

In this study, we investigated how student demographics and performance within online community college courses influenced their participation in a learnersourcing activity that involves generating a multiple-choice question. We found that 37 of the 64 students across the three courses participated in at least one of the learnersourcing activities; these students came from a variety of demographic backgrounds, expressed in terms of self-reported gender, ethnicity, and first-gen status. Our analysis revealed a correlation between the completion of formative assessments and the likelihood of students participating and contributing to a higher number of learnersourcing activities. Interestingly, the top 10% of students, as determined by their quiz score averages, did not participate in any of the learnersourcing activities.

We found no significant relationships between the students' demographic background and their participation with the learnersourcing activities. This may in part be due to our students primarily reporting the same gender and ethnicity, thus decreasing the potential diversity of our sample. While we did not identify any significant effects, our data indicates that a majority of the students from all the reported ethnicities, genders, and first-generation status made at least one contribution to an optional learnersourcing task. While we were encouraged to see that students of all backgrounds were participating, learnersourcing research should continue to collect demographic information to ensure all students are being reached by the activities and interventions. A core benefit of learnersourcing student-generated questions is that their unique perspectives and backgrounds can be incorporated into the questions they create, ultimately avoiding expert-blindspot and contributing to a more diverse pool of questions [171]. However, if the learnersourcing activities are skipped by students, knowing why they are not participating in them and the backgrounds of those students, could potentially inform methods on how to better include all students.

As expected, due to prior research in the area, student participation with the formative assessments in the course was positively correlated with their performance on the summative assessments [34]. We found that students who did more of the formative assessment were also more likely to participate in the learnersourcing activities. There was also a strong positive correlation between the number of formative assessments done and the number of learnersourcing activities students completed. This further suggests students might follow a completionist approach when working through the online materials and not skip the learnersourcing activity, which has been previously reported by [213].

In addition, student performance on both the formative and summative assessments was found to correlate with participation and the number of learnersourcing activities completed. These results indicated that the highest performing students were skipping the task. As mentioned, 7 of the 64 students achieved a perfect score on all the quizzes in their respective courses, yet among these seven students, only one participated in the learnersourcing activities, doing all four offered in their advanced chemistry course. This brings into question if the optional presentation of the learnersourcing activities could be potentially excluding the lower performing students that might benefit the most from these interventions, as well as the top performing ones. While we seek to ideally find a middle ground and engage the full range of learners in the current study, such activities may potentially exclude both the most and least in-need students. The MCQs generated from these top students might be closer to the level of instructor ones due to the advanced domain knowledge they possess [150].

7.5 Limitations

Our contributions should be interpreted against the following limitations. The three community college courses used in this study feature students from three different self-reported ethnicities. While this is representative of the institution-wide demographics, courses at other community colleges might yield a different student population. Additionally, we focused our analysis on data from three STEM courses. Extending this research to more courses from other domains, including non-STEM ones, might provide a more representative sample of students. However, since previous learnersourcing work neglects to provide demographic information, our current focus provides a first step at investigating how the different student populations of a course might be contributing to learnersourcing tasks. Additionally, we did not ask the students to report their native language, which might influence students' willingness to participate in the MCQ generation process.

7.6 Conclusion

In this work, we investigated the optional participation of students in the form of learnersourcing, where they generated multiple-choice questions relevant to the course content. Across three community college courses, our results showed that student demographics had no significant effects on their participation with the learnersourcing activities. However, we had moderate participation from a wide range of students on the task across all courses. Our analysis suggests that students' likelihood of participation with a learnersourcing activity is more dependent on their participation and performance with the other assessments found in the course, rather than on their demographic background. Additionally, we identified several features of student performance in the courses that influenced their participation with learnersourcing activities. These findings demonstrated that better performing students were likely to participate in learnersourcing, yet students at the lowest and highest end of the performance spectrum may still neglect such activities. This work contributed the first study which explicitly investigates the demographics of students participating in learnersourcing activities. It demonstrates that optional learnersourcing activities can still garner participation from a diverse set of students. Future learnersourcing efforts may incorporate participation and performance analytics to encourage students to contribute to learnersourcing tasks.

Chapter 8
Crowdsourcing the Evaluation of Assessments

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Ellen Fang, Huy A. Nguyen, and John Stamper. "Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom's Taxonomy." In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pp. 25-34. 2023.

8.1 Introduction

Large scale learning environments, such as massive open online courses (MOOCs) and other digital courseware platforms commonly utilize multiple-choice questions (MCQs) to measure student learning [36]. These assessments provide beneficial data on student learning, while maintaining objectivity and efficiency in grading. Traditionally, MCQs are authored by a party that has expertise in the given domain, such as an instructor or subject matter expert [32]. However, a continually growing research effort has led to the advancement of MCQ authoring methods that do not rely on experts [92]. For instance, automatic question generation (AQG) systems that leverage the latest techniques in machine learning and natural language processing have allowed MCQs to be created at scale [122]. Keeping the human in the loop, methods such as learnersourcing, which involves students within a course generating novel content to be used by future learners, have also been leveraged to author MCQs at scale [107, 168, 212].

These popular methods allow for the scaling of educational MCQ creation, but they are highly susceptible to generating questions that contain detrimental flaws [4, 98]. Previous work leveraging AQG or learnersourcing methods to create MCQs often utilize the questions without fully assessing their quality or have other students assess the quality [2, 54]. While previous research has demonstrated these methods can be capable of generating expert-level MCQs, the criteria used to judge this quality is often ill described or lacking the pedagogical implications of the questions [10]. For instance, MCQs generated via AQG systems are commonly evaluated using machine learning readability metrics, which commonly omit flaws in the question identified by expert evaluators [206]. In an educational context, when these questions that contain flaws are utilized by students, it can be detrimental to their learning, mislead learning analytics, and ultimately waste valuable student time [46]. Poor question quality can have a detrimental impact on learners in both formative and summative assessments, highlighting the importance of leveraging high quality MCQs effectively in both types of assessments. Evaluating the quality of MCQs before students utilize them is a challenging task that can be difficult to scale, as it often requires human expertise or the time-consuming task of applying a rubric [63].

An emerging area that has the potential to provide the human resources needed for scaling MCQ evaluation is crowdsourcing. Naturally, the challenge with this approach is that the population of crowdworkers is highly varied in their education level and domain knowledge proficiency [161, 164]. Therefore, as a first step towards examining and promoting the feasibility of crowdsourced MCQ evaluation, we studied how crowdworkers can leverage the item-writing flaws (IWF) rubric to assess the quality of MCQs used in formative assessments. The IWF rubric consists of 15 items that assess whether an educational MCQ is acceptable for use in the classroom or not [31, 196]. It provides a standardized way to evaluate the quality of MCQs that includes the pedagogical value of the question and its answer choices through the various criteria. This rubric has previously been applied to educational MCQs used in both formative and summative assessment environments across a plethora of domains [162, 181].

In this work, we explored how crowdsourcing could be leveraged in the quality evaluation of MCQs from the domains of calculus and chemistry. We deployed a crowdsourcing task that had crowdworkers apply the IWF rubric to multiple questions in order to evaluate their quality with respect to their pedagogical value. They also evaluated a different set of questions for their cognitive level, according to Bloom's Revised Taxonomy [119]. Using the wisdom of the crowds, we evaluated if the majority response aligned with expert evaluation of the same questions. This study contributes to the literature on question evaluation and educational crowdsourcing. First, it introduces a method for scaling the evaluation of educational multiple-choice questions. Second, we demonstrate the effectiveness of crowdsourcing the quality assessment of multiple-choice questions. Third, we highlight the domain differences that may impact question evaluation, which has implications on designing and leveraging crowdsourcing in educational tasks.

8.2 Methods

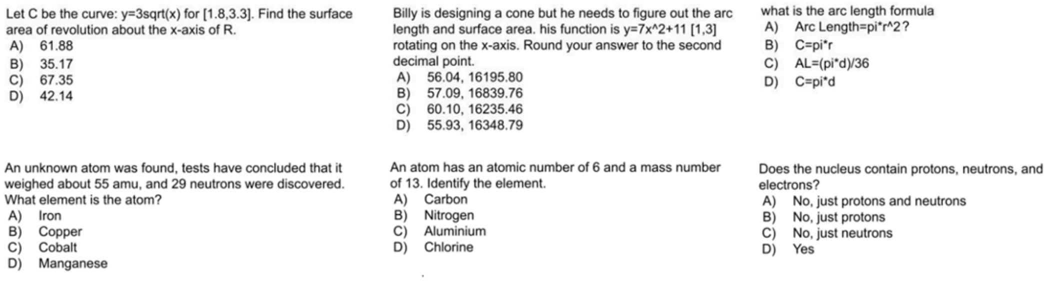
Our study consists of two experiments with the same procedure but involve different domain knowledge. The first domain is calculus, with a focus on the concept and formula of arc length; the second is chemistry, with a focus on atomic theory. In both domains, we deployed an experiment using Amazon's Mechanical Turk (AMT), a general marketplace to crowdsource tasks [178]. Forty crowdworkers on AMT completed the calculus experiment and forty different crowdworkers completed the chemistry experiment, for a total of 80 unique participants. Participants were recruited for the task without using any specific strategy or filters. Instead, the task was posted on the AMT platform, accompanied by a title and description that informed participants they would evaluate multiple-choice questions in one of the two respective domains. In each domain, the tasks took roughly eight minutes to complete. Participants were compensated $1.50 upon finishing, providing a mean hourly wage of $11.25.

The study begins by explaining how multiple-choice questions used in an educational context can target different cognitive processes, such as recall or application. The language used in this description is intended for an audience that does not have a background in learning sciences and we avoided the use of any jargon or other domain-specific terms. Following this, two examples of MCQs that assess at the recall level and two MCQs that assess at the application level of Bloom's Revised Taxonomy are shown to the crowdworker. The content of these questions depends on the domain of the task, such that a crowdworker doing the task for calculus would see example calculus questions. Each example has an accompanying explanation of why it is considered to evaluate this specific level of cognitive ability. Following these instructions and examples, the crowdworker is then presented with three MCQs from their survey's domain, either calculus or chemistry. These three MCQs contain the question text, referred to as the question stem, the correct answer choice, and three alternative answer choices, sometimes referred to as distractors. The crowdworker is then asked to indicate if the question assesses at the recall or application cognitive level. To encourage them to think deeply about their choice, we also asked them to explain why they made their selection. This is a common crowdsourcing tactic that previous research has shown to increase the quality of crowdworker responses [52].

Following this, they advance to the main task of the study, which involves the crowdworker applying the 15 criteria IWF rubric to three separate questions from the task's domain. These rubric criteria are slightly modified to be presented to the crowdworker as a series of yes or no questions, asking if the given MCQ violates the criteria or not. Once all 15 criteria have been applied to the MCQ, they were prompted to briefly explain any flaws they identified in the question text or answer choices. They were also prompted to select if the MCQ they had just evaluated assessed the recall or application cognitive level. After this, they continue to the second and third questions where they repeat the process, evaluating a total of three MCQs in either calculus or chemistry.

8.2.1 Calculus & Chemistry Questions

Each crowdsourcing task utilized a total of six unique questions, with the calculus MCQs assessing the concepts of arc length and the chemistry ones assessing the concepts of atomic theory. The three MCQs used for the initial task of identifying the cognitive process as being recall or application were different from the three MCQs used for the IWF rubric evaluation. All the questions were previously used in an online higher-ed course, either calculus 1 or introductory chemistry, used by several community colleges in the western United States. Figure 8.1 shows the three MCQs used for the calculus task on the top and the three MCQs used for the chemistry task on the bottom. These questions were selected as they contain a differing number of flaws, as well as different types of flaws.



Let C be the curve: y=3sqrt(x) for [1.8,3.3]. Find the surface area of revolution about the x-axis of R.
A) 61.88
B) 35.17
C) 67.35
D) 42.14

Billy is designing a cone but he needs to figure out the arc length and surface area. his function is y=7x^2+11 [1,3] rotating on the x-axis. Round your answer to the second decimal point.
A) 56.04, 16195.80
B) 57.09, 16839.76
C) 60.10, 16235.46
D) 55.93, 16348.79

what is the arc length formula
A) Arc Length=pi*r^2?
B) C=pi*r
C) AL=(pi*d)/36
D) C=pi*d

An unknown atom was found, tests have concluded that it weighed about 55 amu, and 29 neutrons were discovered. What element is the atom?
A) Iron
B) Copper
C) Cobalt
D) Manganese

An atom has an atomic number of 6 and a mass number of 13. Identify the element.
A) Carbon
B) Nitrogen
C) Aluminium
D) Chlorine

Does the nucleus contain protons, neutrons, and electrons?
A) No, just protons and neutrons
B) No, just protons
C) No, just neutrons
D) Yes

Figure 8.1: The three MCQs on the top row were used for the IWF task in the calculus domain and the bottom three MCQs were used for the IWF task in the chemistry domain

8.2.2 IWF Rubric & Cognitive Level

To evaluate the quality of the MCQs used in this study, a set of guidelines to identify item-writing flaws (IWF) in MCQs was utilized. These guidelines come from previous research that established a taxonomy of 31 validated MCQ writing guidelines [86]. The modified version of the rubric used in our study consisted of 15 unique criteria that have been previously tested and validated in prior studies [31, 53, 181]. A complete list of the 15 criteria that make up the rubric can be found in Table 8.1. Note that the criteria span a variety of criteria that assess the different parts of the question, such as the question text, answer choices, and correct option. In addition to evaluating the presence of IWFs, the cognitive process an MCQ assesses was evaluated. Each MCQ was categorized into one of two levels of cognition: recall or application, based on Bloom's Revised Taxonomy, inline with previous research [17, 93]. Recall questions only test the recall of facts or basic comprehension, while application questions assess higher cognitive abilities including the application and analysis of learned concepts.

Item-Writing Flaw	Attributes of questions that do not contain the flaw
Grammatical cues	All options should be grammatically consistent with the stem and should be parallel in style and form
Logical cues	Avoid clues in the stem and the correct option that can help the test-wise student to identify the correct option

Word repeats	Avoid similarly worded stems and correct responses or words repeated in the stem and correct response
Greater detail in the correct option	Often the correct option is longer and includes more detailed information, which clues students to this option
Lost sequence in data	All options should be arranged in chronological or numerical order
Absolute terms	Avoid the use of absolute terms (e.g. never, always, all) in the options as students are aware that they are almost always false
Vague terms	Avoid the use of vague terms (e.g. frequently, occasionally) in the options as there is seldom agreement on their actual meaning
Negative stem	Negatively worded stems are less likely to measure important learning outcomes and can confuse students
Implausible distractors	Make all distractors plausible as good items depend on having effective distractors
Unfocused stem	The stem should present a clear and focused question that can be understood and answered without looking at the options
No correct answer or > 1 correct answer	In single best-answer form, questions should have 1, and only 1, best answer
Unnecessary information in stem	Avoid unnecessary information in the stem that is not required to answer the question
'All of the above'	Avoid all of the above options as students can guess correct responses based on partial information
'None of the above'	Avoid none of the above as it only really measures students' ability to detect incorrect answers
'Fill in the blank'	Avoid omitting words in the middle of the stem that students must insert from the options provided

Table 8.1: The rubric of 15 item-writing flaws used to evaluate the multiple-choice questions

Two evaluators rated each of the six MCQs used in the IWF crowdsourcing task based on the 15 IWF guidelines, using the exact same rubric that the crowdworkers utilized for this study. Both evaluators were experts in the content areas of calculus and chemistry, had extensive experience creating MCQs, and had received multiple training sessions in crafting high-quality assessments. Using the IWF rubric, the evaluators applied the criteria to the text of each question and its 4 answer options. The inter-rater reliability (IRR) values across all six MCQs were calculated between the two evaluators. It includes the percentage agreement and Cohen's Kappa κ statistic [11] as a measure of IRR for all rubric items. The two item raters achieved perfect agreement with one another (100%, κ = 1.00) and there were no discrepancies to resolve for any of the IWF criteria. Although both evaluators were experts with perfect inter-rater reliability, their prior knowledge and linguistic preferences may still influence their application of the IWF rubric.

8.2.3 Data Analysis

After the two experts evaluated the quality of the MCQs using the IWF rubric and the cognitive level they assess, we analyzed the results between them and the crowdsourced application of the rubric. In order to determine if the crowdworkers could effectively apply the IWF rubric for each criteria, we used the majority response to that criteria. For instance, if thirty of the forty crowdworkers in the calculus task said the question violated the first IWF criteria, then we use that as the crowds' response since it is from the majority. This is known as the wisdom of the crowd and is a popular method used to aggregate crowdsourced responses [120].

8.3 Results

8.3.1 IWF Rubric Accuracy

Across all three questions used in the calculus domain, the majority crowdworker vote matched the experts' evaluation perfectly. For the three questions in the chemistry domain, the majority crowdworker vote matched the expert's evaluation for all but one of the criteria for a single question. This criteria the crowdworkers failed to identify was the *logical cue* contained in Q6. Crowdworkers' evaluation of the questions in the calculus domain matched on average 33.40 out of 45 (74.22%) of the IWF criteria identified by expert evaluation. For the chemistry domain, the average was extremely similar, as on average crowdworkers' matched the expert evaluation for 33.43 out of 45 (74.29%) of the IWF criteria.

A breakdown of the crowdworker and expert agreement percentages for each IWF criteria across each question can be found in Table 8.2. Across all six questions, the criteria of *grammatical cues*, *negative stem*, and *unfocused stem* were the three that had the highest average agreement between crowdworkers and the expert evaluators. The three criteria across all six questions with the lowest average agreement, but still in the majority, were *word repeats*, *lost sequence in data*, and *absolute terms*.

Criteria	Chemistry			Calculus		
	Q4	*Q5*	*Q6*	*Q4*	*Q5*	*Q6*
Grammatical cues	90	92.5	77.5	90	80	70
Logical cues	85	87.5	27.5	80	80	67.5
Word repeats	60	72.5	52.5	65	67.5	82.5
Greater detail in the correct option	70	75	65	70	67.5	85
Lost sequence in data	77.5	72.5	62.5	57.5	57.5	60
Absolute terms	72.5	67.5	60	65	70	67.5
Vague terms	82.5	77.5	75	77.5	77.5	77.5
Negative stem	82.5	77.5	80	75	77.5	77.5
Implausible distractors	72.5	77.5	80	82.5	70	57.5
Unfocused stem	77.5	85	82.5	80	85	82.5
No correct answer or > 1 correct answer	85	62.5	72.5	87.5	77.5	70
Unnecessary information in stem	67.5	77.5	72.5	77.5	55	67.5
‘All of the above’	80	77.5	77.5	77.5	77.5	77.5
‘None of the above’	80	82.5	80	80	82.5	75
‘Fill in the blank’	70	70	72.5	77.5	75	80
Average	76.8	77	69.2	76.2	73.3	73.2

Table 8.2: The percentage of crowdworkers that evaluated each IWF criteria the same as the expert evaluators for the given question

The overall agreement between crowdworker and expert evaluations for all three questions in each domain was similar, ranging from 69% to 77%. In chemistry, the top three criteria with the highest agreement were the same for two criteria, but differed on one as chemistry’s third highest criteria was *fill in the blank* instead of *negative stem*. Additionally, crowdworkers had more difficulty with *greater detail* *in the correct option* compared to *lost sequence in data*, for this domain. In calculus, one criteria from the top three lowest and top three highest agreement differed from the overall ones. Crowdworkers struggled more with *all of the above* instead of *word repeats* and did better at identifying *fill in the blank* compared to *logical cues.*

8.3.2 Cognitive Level Accuracy

The average number of questions the crowdworkers correctly identified the cognitive levels of can be seen in Table 8.3. Across the six questions from each domain, the majority of crowdworkers correctly identified the cognitive level for all six calculus questions. For chemistry, five of the six questions had their cognitive level correctly identified by a majority of the crowdworkers. An unpaired two tailed t-test showed there was a strong significant difference in the crowdworker accuracy for identifying the cognitive level of questions in the domain of calculus (M=4.85, SD=2.59) compared to chemistry (M=3.9, SD=0.81), t(39) = 3.257, p < .001.

Calculus Question (Cognitive Level)	Average Accuracy	Chemistry Question (Cognitive Level)	Average Accuracy
1 (application)	82.5%	1 (application)	75%
2 (recall)	82.5%	2 (recall)	90%
3 (application)	80%	3 (recall)	82.5%
4 (application)	75%	4 (application)	40%
5 (application)	80%	5 (recall)	85%
6 (recall)	85%	6 (recall)	85%
Average	80.83%	Average	76.25%

Table 8.3: The average accuracy of crowdworkers in identifying the cognitive level of questions in each domain, with questions 1-3 being in the pretest and questions 4-6 being used in the IWF task

The cognitive level identification task was split into two sections. In the first section, the crowdworkers were asked to determine the cognitive level of three questions as part of a pretest at the beginning of the survey. In the second section, they were instructed to identify the cognitive level of each question immediately after applying the IWF rubric to it. We hypothesized that crowdworkers would be more accurate on the questions they applied to the IWF rubric to, since they spent more time on task with those questions. However, the results from the three calculus questions from the pre-test compared to the three calculus questions in the IWF task indicate there was no significant difference in the cognitive level identification accuracy, t(39) = - 0.529, p = .599. Similar results were found for chemistry, as there was no significant difference observed between the accuracy on the three pretest questions and the three IWF task questions, t(39) = 1.817, p = .077.

We further hypothesized that crowdworkers who performed better at the cognitive level identification task would also perform better when applying the IWF rubric. For calculus, there was a strong significant and positive correlation between a crowdworker's accuracy on the cognitive level task and their accuracy on the IWF task, r(39) = .60, p < .005. Similar results were observed for chemistry, as there was also a strong significant and positive correlation between the accuracy of the cognitive level and IWF identification tasks, r(39) = .48, p < .005. Additionally, we found no significant difference between the number of flaws identified in a question with the cognitive level it assesses in this study.

8.4 Discussion

In this study, we investigated the feasibility of crowdsourcing the evaluation of educational multiple-choice questions (MCQs). We found that in the domain of calculus, the crowdsourced application of the IWF rubric to three MCQs matched the expert application of the rubric exactly. In the domain of chemistry, we found similar results between the crowdsourced task and expert evaluation, achieving an exact match on every criteria except one. On average, crowdworkers matched 74% of the 15 IWF criteria applied across all three questions in both domains. For identifying the cognitive level each question assesses, crowdworkers correctly identified it for all six calculus questions and five of the six chemistry questions. Our results showed that crowdworkers with little to no domain expertise can accurately evaluate the quality of MCQs from higher-ed STEM domains by applying the IWF rubric.

When applying the IWF rubric to six MCQs - three from calculus and three from chemistry - the crowdworkers consistently demonstrated high accuracy in evaluating three specific criteria. These criteria were *grammatical cues*, *negative stem*, and *unfocused stem*. All three of these flaws were not present in the MCQs from either domain, which a majority of the crowdworkers correctly identified. Two of these criteria involve surface level features of the question, such as the grammar or use of a negative word in the question's text. These criteria could be evaluated using automatic methods through implementation of a natural language processing library or even keyword matching [218]. However, identifying that a question stem is unfocused, causing it to be misunderstood or unanswerable without looking at the answer choices, would be more challenging to programmatically assess, as it may rely on prior knowledge and a more comprehensive understanding of language.

While the crowdsourced majority applied the IWF rubric perfectly to the calculus MCQs, they missed a single criteria present in the last chemistry question. This criteria is referred to as *logical cue*, which asked the crowdworkers "*Are the question text and correct answer choice free of any clues that may help identify the correct answer?*". For this question, viewable in the bottom right of Figure 8.1, it may appear at first that there are no cues that indicate the correction option. However, there is a convergence cue present in the question, as the words *protons* and *neutrons* are each repeated twice throughout other answer choices, suggesting that the correct option might be a combination of the two. While rare, these convergence cues can be found in multiple-choice questions, as the alternative answer options tend to share keywords used in the correct answer [234]. A previous study by [221] analyzed 2,770 MCQs from medical exams administered at their university and found that 0.2% of them contained this flaw.

In this study, forty unique crowdworkers were employed to evaluate chemistry and calculus questions separately. This sample size was chosen based on previous crowdsourcing studies that utilized user evaluation to achieve consensus, determining that forty crowdworkers provided saturation [161]. Additionally, the agreement threshold of 50% or higher with the expert evaluation aligned with prior crowdsourcing research [120]. It was observed that using a smaller number of crowdworkers would yield different results, as the majority did not immediately match the expert evaluation for all criteria. The ultimate goal is to identify consensus or a clear majority while minimizing the number of crowdworkers, thus saving time and money. However, it is important to note that this optimal cutoff may vary depending on the crowdworkers and the type of questions, necessitating further research in this area in the future.

Crowdworkers correctly identified the cognitive level of all six MCQs used in the calculus task and five of the six MCQs in the chemistry task. While this is a high accuracy rate for a task that can be challenging to even experts [8], in this case the crowdworkers had a 50% chance to correctly guess the answer. When prompted to identify the cognitive level of a given question, they were only presented with the two options of *recall* or *application*. We intentionally designed it to include just these two options, one from the lower levels of Bloom's Revised Taxonomy and one that represents a higher order question [17]. For this study, we wanted to see if crowdworkers could make this distinction of lower or higher cognitive process before asking them to select from all six levels of the taxonomy. Previous research often questions the validity of all six levels of the taxonomy, as it may create the misconception that cognitive processes at each level are separate and that certain skills are more challenging or significant than others [188]. However, previous research has validated the distinction between lower and higher order cognitive processes, although it is not necessarily aligned with the specific six levels of Bloom's Revised Taxonomy [187].

The chemistry question that crowdworkers misidentified the cognitive level of can be found in the bottom left of Figure 8.1. We believe crowdworkers incorrectly thought this was a recall question since the answer choices only contain the name of elements on the periodic table. However, to identify the correct element, the student needs to use the two provided values in the question to make a calculation. This makes the question at the *application* level, as you need to apply a particular equation to achieve the correct answer.

Finally, we found a strong significant difference between crowdworker accuracy on the IWF portion of the task based on their accuracy of cognitive levels. We attribute this to potentially identifying crowdworkers that were devoting the most effort and paying attention to the task, rather than those crowdworkers having prior knowledge about Bloom's Revised Taxonomy or the IWF rubric. Interestingly, across both calculus and chemistry, there was no significant difference in the crowdworker accuracy for identifying the cognitive level of the MCQs they applied the IWF rubric on. We believed since crowdworkers were spending more time on those questions, as they applied the 15 rubric criteria to them, that they would have a better understanding of what it is asking and thus achieve a higher accuracy. However, this was not the case for the present study, as no significant difference was found.

8.5 Limitations & Future Work

We identified several limitations in the present study that might influence the results in other domains or with other questions. For this work, we only utilized questions from the two STEM domains of calculus and chemistry that were used in higher-ed courses. Including questions from other domains and from different grade levels would likely alter the outcome of this task. Secondly, depending on when the study is deployed, the pool of crowdworkers that complete the task might be better or worse. Even with demographic surveys at the start of the task, it can be difficult to truly understand the backgrounds of the crowdworkers and how it might influence their success or failure for evaluating these MCQs. Additionally, we have a limited sample size of questions that assess two different cognitive levels. Our limited sample is constrained by a set of questions for which we have multiple expert evaluations using the IWF rubric. Finally, only two levels of Bloom's Revised Taxonomy were used in this study. While these two levels were selected due to them denoting lower order (recall) or higher order (application) cognitive levels based on prior work [93], participants could have potentially correctly guessed between the two options when answering those questions.

Future work should look to expand the crowdsourcing of educational MCQ evaluation using other domains and different questions. While the domains we used are fairly complex, different domains might be more or less suitable for this crowdsourcing task. The Bloom's Revised Taxonomy levels used could also be expanded to include all six classifications, rather than only using recall and application. To help scale the evaluation of MCQs using the IWF rubric, some of the criteria could be automatically assessed using programmatic methods. For instance, using string matching one could easily identify if a question contains *all of the above* or is a *fill in the blank* question. This in turn could make the evaluation process more efficient, by requiring the crowdworkers to only evaluate the MCQs using criteria that require human knowledge. Another potential that builds on this work is having the crowdworkers suggest or make improvements to the MCQs based on the flaws that they identified. This could help yield more high quality questions, as sometimes MCQs contain one or two flaws that are trivial to fix, which could then make them into high quality questions.

8.6 Conclusion

In this paper, we proposed a novel crowdsourcing task for evaluating the quality of educational multiple-choice questions using criteria from the item-writing flaws rubric. The results indicate that crowdworkers can accurately assess the quality of multiple-choice questions across distinct subject areas. We highlight how certain flaws may be easier or harder for crowdworkers to identify, depending on the subject area. Our results also demonstrate how crowdworkers can effectively identify the cognitive level of questions at the lower and higher levels of Bloom's Revised Taxonomy. These results provide the demonstrated success of a method for scaling the evaluation of educational MCQs. This work also opens up further opportunities for developing scalable methods for evaluating educational questions using features related to their pedagogical values.

Chapter 9
Automatically Evaluating MCQs for Quality

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Huy A. Nguyen, Tianying Chen, and John Stamper. "Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods." In *European Conference on Technology Enhanced Learning*, pp. 229-245. Cham: Springer Nature Switzerland, 2023.

9.1 Introduction

Multiple-choice questions (MCQs) are a widely used form of assessment in higher education, both for formative and summative evaluations. MCQs are advantageous because of their efficiency to score, objective grading, ability to generate item-analysis data, and the shorter time required for students to respond [36]. In recent years, the task of authoring educational MCQs is no longer specific to instructors, and the popularization of automatic question generation (AQG) systems further scaled up this process [122]. Another method for scaling the creation of educational MCQs is having students take part in the process of question creation, commonly referred to as a form of learnersourcing [212]. Student-generated questions often have higher quality, and target more complex cognitive processes compared to AQG [91]. The process of generating questions also has educational benefits for students and can lead to positive learning outcomes, such as improved retention and transfer [107].

While student-generated questions typically have higher quality than those created through automated methods, their quality may widely vary due to multiple uncontrollable factors [162]. On one hand, poorly designed MCQs may exhibit characteristics that can be exploited by pattern recognition and guessing, thus leading to shallow learning [69]. On the other hand, ensuring high-quality MCQs, whether created by AQG or students, is itself a challenging task. Common evaluation methods used by previous research include using experts, other students, or automated methods [140]. Even though automatic methods are most efficient, they come with important caveats. Notably, existing automated methods often rely on the surface-level features of a question, such as the readability of text length, without considering the pedagogical value it adds to the assessment [11]. Additionally, these methods are often applied to datasets consisting of questions targeted at lower academic grade levels, such as basic reading comprehension, or questions that are not used in an educational context at all [91]. While the use of human experts might provide the most accurate assessment of question quality, the manual evaluation process often lacks standardization and efficiency [122]. However, human evaluation can provide a more in-depth analysis, considering the question's potential to support learning. For instance, the Item-Writing Flaws (IWFs) rubric is an effective evaluation method which considers the pedagogical value of the question and its answer choices through various criteria [31, 53]. This rubric typically consists of 19 items that assess whether an educational MCQ is acceptable for use in the classroom or not. However, applying this rubric to a large number of questions can be time-consuming and often requires human expertise.

To address this gap, we explored two automatic methods to evaluate educational MCQs using the IWF rubric. The first method utilizes a rule-based approach to apply the rubric, making it easy to modify and maintain interpretability, while not requiring a large training dataset. Our second method relies on GPT-4, a large multimodal model capable of processing text inputs and producing text outputs, which has achieved human level performance on various professional and academic benchmarks [175]. This second method prompts GPT-4 to apply the IWF criteria to the provided questions one at a time. Using student-generated questions from four distinct subject areas, we evaluated both methods and compared them to human expert evaluation that also utilized the IWF rubric. We investigate to what extent a rule-based multi-label classifier and GPT-4 can accurately identify IWFs in student-generated MCQs.

9.2 Methods

9.2.1 Dataset

The datasets used in this study were collected from a digital learning platform used by several public universities and community colleges in the western United States. The data comes from students using the platform in their respective courses during the 2020 and 2021 academic years. The four courses are introductory Chemistry, introductory Biochemistry, introductory Statistics, and a course on learning how to effectively collaborate, referred to as CollabU. Students in these courses were undergraduates, towards the beginning of their studies, and pursuing either a two- or four-year degree.

As students worked through the digital learning materials on the platform in their respective courses, they were prompted to create a multiple-choice question (MCQ). The prompt asked students to create a single MCQ about a topic they had recently learned in their course. Each MCQ consists of a question text, known as the stem, and four answer choices, one of which must be denoted as correct. The creation of this MCQ was done directly in the digital learning platform with no additional tools utilized. Students did not receive any assistance or feedback as they created their questions. Additionally, it was presented in the same visual manner as the other activities found on the platform. From each of these four courses, we randomly selected 50 student-generated questions to utilize for this study, resulting in a total of 200 MCQs.

9.2.2 Human Evaluation

In order to assess the quality of the student-generated MCQs, we utilized a series of guidelines for identifying Item-Writing Flaws (IWFs), which are based on a taxonomy of 31 multiple-choice item-writing guidelines [85]. The exact rubric we used for the study was a modified version that consists of 19 unique items and has been used and validated in previous studies [31, 53, 162]. Following [221], a question with 0 or 1 flaw identified by the rubric is considered acceptable and any questions with 2 or more flaws is considered unacceptable. This distinction is used to determine if a question could be utilized in a class as a formative assessment that the instructor would trust. A full description of the 19 items that make up the rubric can be found in Table 9.1.

Item-Writing Flaw	**Attributes of questions that do not contain the flaw**
Ambiguous or unclear information (87.50%, κ = 0.66)	Questions and all options should be written in clear, unambiguous language
Implausible distractors (96.00%, κ = 0.82)	Make all distractors plausible as good items depend on having effective distractors
None of the above (100%, κ = 1.00)	Avoid none of the above as it only really measures students ability to detect incorrect answers
Longest option correct (98.50%, κ = 0.83)	Often the correct option is longer and includes more detailed information, which clues students to this option
Gratuitous information (89.50%, κ = 0.71)	Avoid unnecessary information in the stem that is not required to answer the question
True/false question (100%, κ = 1.00)	The options should not be a series of true/false statements
Convergence cues (89.50%, κ = 0.70)	Avoid convergence cues in options where there are different combinations of multiple components to the answer
Logical cues (88.00%, κ = 0.68)	Avoid clues in the stem and the correct option that can help the test-wise student to identify the correct option
All of the above (100%, κ = 1.00)	Avoid all of the above options as students can guess correct responses based on partial information

Fill-in-blank (100%, κ = 1.00)	Avoid omitting words in the middle of the stem that students must insert from the options provided
Absolute terms (100%, κ = 1.00)	Avoid the use of absolute terms (e.g. never, always, all) in the options as students are aware that they are almost always false
Word repeats (97.00%, κ = 0.83)	Avoid similarly worded stems and correct responses or words repeated in the stem and correct response
Unfocused stem (94.50%, κ = 0.79)	The stem should present a clear and focused question that can be understood and answered without looking at the options
Complex or K-type (94.00%, κ = 0.78)	Avoid questions that have a range of correct responses, that ask students to select from a number of possible combinations of the responses
Grammatical cues (92.50%, κ = 0.76)	All options should be grammatically consistent with the stem and should be parallel in style and form
Lost sequence (97.00%, κ = 0.89)	All options should be arranged in chronological or numerical order
Vague terms (98.50%, κ = 0.93)	Avoid the use of vague terms (e.g. frequently, occasionally) in the options as there is seldom agreement on their actual meaning
More than one correct (100%, κ = 1.00)	In single best-answer form, questions should have 1, and only 1, best answer
Negative worded (100%, κ = 1.00)	Negatively worded stems are less likely to measure important learning outcomes and can confuse students

Table 9.1: The rubric of 19 Item-Writing Flaws used to evaluate the student-generated multiple-choice questions. The bracketed numbers indicate agreement percentage between raters and Cohen's κ value for each item

Two item raters evaluated each student-generated MCQ, following the 19 IWF guidelines. Both raters had content-area expertise across all four domains, ample experience developing multiple-choice questions, and multiple prior training sessions in writing high quality assessments. Using the IWF rubric, the raters went through each of the 200 student-generated MCQs and applied the rubric to the question text and accompanying answer choices for each student contribution. The inter-rater reliability (IRR) values between the two evaluators for each rubric item are also reported in Table 9.1. It includes the percentage agreement and Cohen's Kappa κ statistic [148] as a measure of IRR for all rubric items. All items were at either a near perfect or substantial level of agreement between the raters. The two evaluators then met to resolve any disagreements in their evaluations and discussed discordant questions until they reached a consensus on the coding. We acknowledge that, despite the two expert evaluators' backgrounds and high IRR, they could still interpret the student-generated questions differently, based on their prior knowledge and linguistic preferences [11].

9.2.3 Rule-Based Evaluation

The task of automatically applying the Item-Writing Flaws rubric to MCQs is a multilabel classification problem, as each question may be matched with several criteria [6]. In order to implement this automated method, we followed a rule-based approach that applies each individual rubric criteria via its own logic. Rule-based approaches have been used in similar educational tasks such as classifying the Bloom's Revised Taxonomy of a question [88]. Such an approach is particularly effective when the problem suffers from a lack of training data, as is the case in the present study, due to a lack of public datasets containing questions that are evaluated for their educational quality [95]. Furthermore, a rule-based approach allows for vastly improved interpretability compared to traditional black-box classification approaches, such as neural networks [231].

Working alongside the human evaluators, we constructed a script that is composed of a programmatic method for each of the 19 IWF rubric criteria. It uses several Python libraries and three different pre-trained large language models (LLMs) to implement the 19 different criteria. The logic for many of the criteria involved string manipulation, such as checking if the longest option was the correct answer. Other criteria involved the use of standard NLP techniques, such as Named Entity Recognition or Part-of-Speech tagging [218] to help identify if a word is repeated between the question's stem and correct answer. The more challenging and advanced criteria, such as identifying if a question contains implausible distractors, involved the use of LLMs. For instance, a RoBERTa classifier pretrained on the Corpus of Linguistic Acceptability (CoLA) was utilized to help identify ambiguous or unclear information in a question's stem [121]. To determine if a question contained more than one correct answer, we leveraged GPT-4's capabilities for question answering [175]. For a more detailed explanation of the programmatically implemented 19 IWFs criteria, the final code is made publicly available[2], however the student question data is currently private and can be made available upon request.

9.2.4 GPT-4 Evaluation

The second automatic evaluation method utilizes GPT-4, a transformer-based multimodal model pre-trained to predict the next token in a document [175]. We utilize GPT-4 as our second automated method as it has achieved human-level performance on academic tasks, such as standardized college-level exams in

[2] https://github.com/StevenJamesMoore/ECTEL23/blob/main/IWF.ipynb

Psychology, History, and Math. It has also achieved state-of-the-art performance on traditional machine learning benchmarks, such as the MMLU, which consists of 57 tasks from a variety of domains that are used to demonstrate a model's extensive world knowledge and problem-solving ability [89]. What also makes GPT-4 unique compared to many other language models, is the ability to follow natural language prompts to perform specific tasks [135]. These prompts serve as instructions for the model to perform, such as providing the model with a rubric and multiple-choice question and then prompting it to apply the rubric criteria to the question.

The exact wording of the prompts provided to GPT-4 can drastically influence the output the model provides [127]. Towards this end, our task involved providing GPT-4 with a single IWF rubric criteria at a time and having it state if the provided question satisfied that given criteria or not. The names of the criteria we used as well as their definitions are nearly identical to the ones shown in Table 9.1. We opted to directly use the prompts rather than continually engineering prompts to determine the best output. While we believe refinement of the prompts is valuable future work, we wanted to see how well GPT-4 would perform applying the same IWF rubric and giving it instructions akin to what would be provided to a human evaluator. Specifically, the prompt we provided GPT4 for each IWF rubric criteria and question states: *Begin your response with yes or no, does this multiple-choice question satisfy the criteria relating to {criteria}: {definition}? Explain why. {question}*. The rubric criteria, definition of the criteria, and the multiple-choice question including all answer options are input into the prompt respectively. Additionally, we utilized the default parameters of the model and accessed it using the GPT-4 API via the Python programming language.

Note, the prompt instructions also asked GPT-4 to provide an explanation as to why a question satisfies or violates the criteria, which was done to encourage a more thorough and accurate response from the model [103]. A human evaluator went through each of the responses and coded them as GPT-4 indicating if the criteria was satisfied or violated. Although we had originally intended to use a simple "Yes" or "No" response to indicate whether the criteria were met, we found that this approach was not always clear in distinguishing whether the criteria had been violated or satisfied.

9.3 Results

9.3.1 Automatic Methods versus Human Identification

The 19 IWF criteria were automatically applied to all 200 student-generated questions, resulting in a total of 3800 classifications. The rule-based method matched 90.87% of human classifications, achieving an exact match ratio of 15%, where all of the 19 IWF criteria matched the human evaluation for the question.

The GPT-4 method matched 78.89% of human classifications, achieving an exact match ratio of 12%. We also considered the Hamming loss, which is a measure of the difference between two sets of binary labels and calculated as the fraction of labels that are incorrectly predicted [225]. The rule-based method achieved a Hamming Loss of 0.09 and the GPT-4 method achieved a Hamming Loss of 0.21, indicating that on average 9% and 21% of the flaws were misclassified respectively. Table 9.2 displays the number of IWFs assigned to questions for each evaluation method grouped by counts. A paired t-test showed a small significant difference in the number of IWFs identified for each question by the human (M = 1.6, SD = 1.3) and rule-based (M = 2.1, SD = 1.4) methods, t(199) = 5.59, p < .001. The rule-based evaluation method more commonly identified potential flaws in the questions compared to the humans. Another paired t-test showed a significant difference in the number of IWFs identified for each question by the human (M = 1.6, SD = 1.3) and GPT-4 (M = 4.2, SD = 3.4) methods, t(199) = 11.8, p < .001. The GPT-4 evaluation identified even more potential flaws in the questions compared to both the human and rule-based methods. While the human and rule-based methods never found more than six IWFs per question, the GPT-4 method found up to thirteen

Number of Flaws	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Human Evaluation	39	72	42	28	9	6	2	0	0	0	0	0	0	0
Rule-based Evaluation	23	49	56	34	27	8	1	0	0	0	0	0	0	0
GPT-4 Evaluation	30	23	25	24	13	17	7	19	11	11	9	7	1	1

Table 9.2: Counts of IWFs per question from all three evaluation methods

When the quality of the questions were labeled as acceptable (< 2 IWFs) or unacceptable (≥ 2 IWFs), a chi-square test revealed there was a significant relationship between the question quality and three evaluation methods, χ2(2,N = 200) = 36.64, p < .001. Between the three methods, GPT-4 was more likely to evaluate a question as having unacceptable quality. The human evaluation identified 111 acceptable and 89 unacceptable questions, while the rule-based evaluation matched 130 (65%) of these (57 acceptable, 73 unacceptable). The GPT-4 evaluation matched 123 (62%) of the human quality evaluations (44 acceptable, 79 unacceptable). Figure 9.1 shows the confusion matrices for the quality classifications based on the number of IWFs found in each question between the human and rule-based evaluation and the human and GPT-4 evaluation.

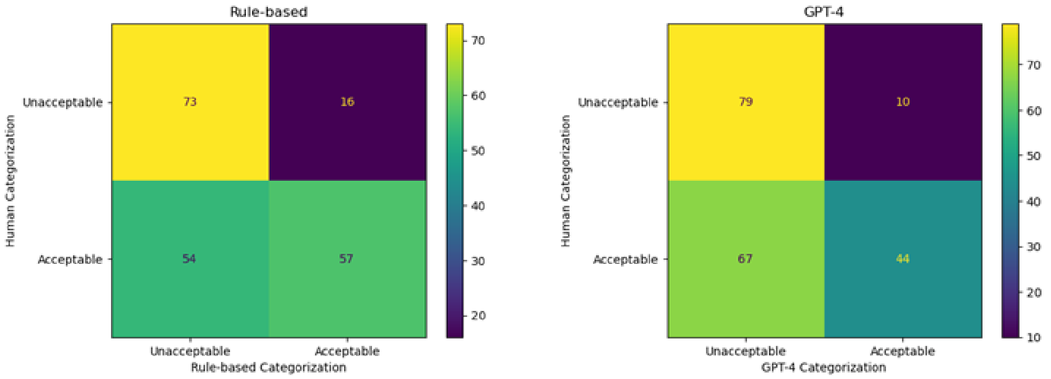


Figure 9.1: Confusion matrices for the classification of a question's quality for the rule-based method (left) and the GPT-4 method (right)

9.3.2 Impact of the Domain

The automatic evaluation methods, rule-based and GPT-4, performed differently across criteria and domains, with the rule-based method outperforming GPT-4 on all four domains. Table 9.3 shows the performance of all three evaluation methods across all four domains. Between the datasets, we use F1 scores to evaluate success. Since a majority of the questions meet the criteria rather than violate them, the F1 score provides a better measure over accuracy, as it includes false negatives and false positives. From Table 9.3, we observe that the rule-based and GPT-4 methods commonly matched human evaluation for some criteria, such as *none of the above* and *negative worded*, and performed poorly for other criteria, such as *logical cues* and *more than one correct*. In particular, the rule-based method achieves high F1 scores for *longest option correct* and *true/false question* compared to GPT-4. Note, the rules for these two criteria can be easily implemented programmatically, as they only check for text length and keywords.

Both the rule-based and GPT-4 methods have a lower micro-average F1 score, the computed proportion of correctly classified observations out of all observations, for the Chemistry and Biochemistry courses compared to Statistics and CollabU. This may be in part due to the similar domains of these science courses, where the human evaluators focused more on the objective of the questions, rather than the grammar, while the automated methods did not. Additionally, some of the poor performance related to F1 scores is due to the small number of that flaw being found in the questions. For instance, *gratuitous information* and *vague terms* have poor performance by F1 score, but those flaws are quite rare across all four courses. Ultimately, the rule-based method outperformed GPT-4, by measure of micro-average F1 score, across all four domains.

Item-Writing Flaws	Chemistry	Biochemistry	Statistics	CollabU

		Hum	Rule	GPT	Hum	Rule	GPT	Hum	Rule	GPT	Hum	Rule	GPT
ambiguous information	N	12	2	11	7	4	25	14	11	25	7	7	40
	F1	-	0.14	0.61	-	0.00	0.25	-	0.40	0.41	-	0.29	0.30
implausible distractors	N	10	6	17	3	9	16	8	9	17	25	21	29
	F1	-	0.12	0.30	-	0.33	0.11	-	0.82	0.16	-	0.83	0.78
none of the above	N	7	8	6	3	3	3	1	1	1	1	1	1
	F1	-	0.80	0.62	-	1.00	0.67	-	1.00	0.00	-	1.00	1.00
longest option correct	N	5	5	6	2	2	11	3	3	6	9	10	13
	F1	-	0.80	0.18	-	1.00	0.00	-	1.00	0.44	-	0.95	0.36
gratuitous information	N	0	0	2	7	5	25	3	0	18	0	0	19
	F1	-	-	0.00	-	0.67	0.38	-	0.00	0.29	-	-	0.00
true/false question	N	2	3	1	6	11	11	1	1	3	3	3	4
	F1	-	0.80	0.67	-	0.59	0.24	-	1.00	0.00	-	1.00	0.29
convergence cues	N	2	12	12	12	36	13	9	11	18	10	16	16
	F1	-	0.14	0.14	-	0.50	0.24	-	0.80	0.30	-	0.77	0.23
logical cues	N	2	2	9	2	6	15	3	1	18	10	0	23
	F1	-	0.00	0.00	-	0.00	0.00	-	0.00	0.10	-	0.00	0.36
all of the above	N	2	0	1	3	0	2	1	0	1	2	1	3
	F1	-	0.00	0.67	-	0.00	0.80	-	0.00	1.00	-	0.67	0.80
fill-in- the-blank	N	2	2	3	4	3	2	0	0	0	1	1	4
	F1	-	1.00	0.00	-	0.86	0.67	-	-	-	-	1.00	0.40
absolute terms	N	2	6	1	6	19	9	0	2	1	7	10	6
	F1	-	0.00	0.00	-	0.40	0.27	-	0.00	0.00	-	0.71	0.62
word repeats	N	0	0	2	8	0	7	0	0	6	3	2	5
	F1	-	-	0.00	-	0.00	0.13	-	-	0.00	-	0.00	0.25
unfocused stem	N	0	2	7	5	4	15	7	3	17	4	4	28
	F1	-	0.00	0.00	-	0.44	0.30	-	0.40	0.50	-	0.75	0.25
complex or K-type	N	2	0	6	6	9	6	3	2	14	1	1	27
	f1	-	0.00	0.25	-	0.53	0.50	-	0.80	0.24	-	1.00	0.07
grammatical cues	N	1	3	17	13	24	14	5	13	20	11	18	36
	F1	-	0.00	0.11	-	0.65	0.30	-	0.44	0.32	-	0.76	0.38
lost sequence	N	0	2	12	2	0	12	11	11	17	0	0	7
	F1	-	0.00	0.00	-	0.00	0.29	-	0.91	0.43	-	-	0.00
vague terms	N	0	0	0	2	0	4	0	0	3	0	0	10
	F1	-	-	-	-	0.00	0.00	-	-	0.00	-	-	0.00
more than one correct	N	0	13	5	0	4	14	0	20	16	4	10	22
	F1	-	0.00	0.00	-	0.00	0.00	-	0.00	0.00	-	0.43	0.15
negative worded	N	0	0	0	8	14	15	2	3	2	6	7	6
	F1	-	-	-	-	0.64	0.70	-	0.80	1.00	-	0.92	1.00
micro-avg		-	0.30	0.25	-	0.48	0.28	-	0.56	0.30	-	0.70	0.36
totals		49	66	118	99	149	219	71	91	203	104	112	299

Table 9.3: The count of flaws (N) and performance (F1) of the human evaluation (Hum) compared to both the rule-based (Rule) and GPT-4 (GPT) methods across all four domains. A dash (-) in the table indicates that the flaw was not present in

any question of that dataset based on human evaluation and therefore no F1 score is computed

9.3.3 Common Item-Writing Flaws

The most frequently identified violated criteria varied across the three methods, although in some domains the rule-based and GPT-4 methods had similar classifications to the human evaluation. Table 9.3 shows that the *implausible distractor* criteria was violated the most across all questions in human evaluation, whereas *vague terms* was the least violated. On the other hand, the rule-based method found *convergence cues* to be the most commonly violated criteria and *vague terms* to also be the least violated. As for the GPT-4 method, the most commonly violated criteria was *ambiguous unclear information*, and *all of the above* was the least violated. Although the most frequently violated criteria varied across all three methods, the rule-based and GPT-4 methods shared similar results with human evaluation. Specifically, the rule-based method's most violated criteria ranked as the third most violated criteria in human evaluation, while the GPT-4 method's most violated criteria ranked second.

A Pearson correlation coefficient showed there was a significant positive correlation between the number of flaws identified for each criteria between the rule-based and human evaluations for Biochemistry (r(17) = .747, p < .001), Statistics (r(17) = .496, p < .05), and CollabU (r(17) = .833, p < .001). However, this correlation was not found to be significant for Chemistry (r(17) = .191, p = .433). A Pearson correlation coefficient was also computed for the number of flaws identified for each criteria between the GPT-4 and human evaluations. There was a significant positive correlation found for Statistics (r(17) = .756, p < .001), and CollabU (r(17) = .4702, p < .05). No significant correlation was found for Chemistry (r(17) = .443, p = .057) and Biochemistry (r(17) = .262, p = .278). This suggests that for Statistics and CollabU, both automated methods identified similar trends in the violated criteria – i.e., if a flaw was commonly found by human evaluation, it was also likely to be commonly found by the automated methods.

9.4 Discussion

In this work, we developed an automatic rule-based method and assessed its performance compared to GPT-4 and human annotation for evaluating the quality of educational MCQs using the IWF rubric. In contrast to prior research, we employed criteria that pertain directly to the pedagogical value of the question across multiple dimensions. We found that this method can perform at a level comparable to human evaluation for certain rubric criteria and outperforms GPT-4 in the same task across all rubric criteria. The rule-based method was effective in evaluating questions across four distinct subject areas, even with the presence of domain-specific jargon. When comparing the results of our automatic evaluation methods to human evaluation, we identified commonly found IWFs in student-generated questions across the four subject areas. Our results suggest that using a rule-based multi-label classification method can achieve a high level of accuracy while also maintaining interpretability, which the GPT-4 method lacks.

Both of the automatic methods' classifications were stricter, in the sense that they assigned many more IWFs to the student-generated questions than human experts, particularly GPT-4. However, this is preferable to being less strict, as guaranteeing high-quality questions during the evaluation process is crucial so as to not disrupt student learning. Additionally, both of the automatic methods could easily help filter out questions whose quality is too low for human review, e.g., if a question has four or more IWFs, it would likely take substantial time to review and could be dropped. This filtering capability of the rule-based method is supported by our results showing that it matches 65% of human classification when categorizing questions as *acceptable* or *unacceptable*, based on the IWF count. This method of binary classification of quality is commonly used in MCQ evaluation and has a performance level comparable to other models using similar educational datasets [174, 195]. Additionally, these automatic methods could be identifying criteria that were missed by the human evaluators, rather than misclassifying questions with the IWF rubric criteria.

While GPT-4's training data included material from the four course domains used in this study, its black-box nature poses challenges in interpreting why it might be misclassifying specific IWF criteria [175]. For instance, the GPT-4 method achieves extremely low F1 scores for *gratuitous information*, *unfocused stem*, and *vague terms*, all of which relate to the question's stem being unnecessarily verbose. Our analysis revealed that GPT-4 identifies a significantly high number of these three flaws across questions in each domain compared to the human and rule-based methods. This could mean that GPT-4 is mistakenly combining these criteria due to their similarity, marking them all as violated based on a single flaw. In contrast, the rule-based method can be designed to implement each criteria explicitly without overlapping with other flaws.

Across the four different subject areas utilized in this study, we found that both the rule-based and GPT-4 methods performed better on Statistics and CollabU, compared to Chemistry and Biochemistry. This may be in part due to the latter two domains containing questions that use more terminologies and jargon, making some of the NLP techniques less effective [47]. Interestingly, the rule-based method achieved more than double the micro-average F1 score of the GPT-4 method in CollabU. GPT-4's worse performance in this case may be due to proper nouns being included in the question text. The human evaluators familiar with the course content would find the usage of the proper nouns acceptable and the rule-based method does not leverage proper nouns in many of the criteria.

However, GPT-4 may identify these as errors in the question as it lacks the necessary context to know if they are essential to the question or not.

Compared to the other course domains, CollabU, a course on learning how to effectively collaborate, may have more recall-level student-generated questions. In contrast, the other domains may include more complex questions that involve formulas or numbers that are challenging to decipher for the automatic methods. Criteria such as *lost sequence* are also more applicable to domains such as Chemistry or Statistics as they may include question options that are purely numerical, causing the arrangement of options to matter. Additionally, both automatic methods performed the worst in Chemistry and Biochemistry, two closely related science courses. IWF criteria such as *grammatical cues* and *convergence cues* were excessively identified by both methods compared to the human evaluators. With the subjectivity that arises from human evaluation, even when applying a standardized rubric, it is possible the evaluators were less focused on grammar in this domain and more focused on the objective of the question, prioritizing *what* was being asked more than *how* it was being asked. This highlights the need for automatic evaluation methods that can focus on both the syntax and the question’s content that is critical to the domain and pedagogy. In contrast to human evaluation, automatic methods scale easier, reduce human subjectivity leading to enhanced replicability, and can be used by individuals without domain expertise.

Finally, the rule-based method demonstrated effectiveness in identifying IWFs that are common in accordance with human evaluation in each of the datasets analyzed. In line with previous research, *ambiguous or unclear information* and *implausible distractors* were two of the most identified flaws across all questions by the human, rule-based, and GPT-4 methods [196, 221]. Our analysis revealed that 50% of the questions in the CollabU dataset exhibited the *implausible distractors* flaw. Again, this may be attributed to the recall-based nature of the material in this domain, which could make it challenging for students to generate plausible alternative options for the questions. In contrast to [196, 221], our datasets contained a high percentage of questions with the *convergence cues* flaw. This might be a result of the digital interface that students used to construct MCQs in our study, as it might have encouraged them to copy the correct answer and then modify it, leading to the prevalence of this flaw. In turn, these findings can inform teachers of the common flaws that they should focus on when refining student-generated MCQs and providing them with feedback on the task.

We expected both automatic methods to perform highly on the IWF criteria *more than one correct*, as they both leverage GPT-4, which has achieved success in these course domains [175]. However, the presence of different flaws, such as incorrect grammar or the inclusion of proper nouns, may cause the question to be confusing, potentially misleading GPT-4 into incorrectly answering some of the questions. Additionally, while criteria such as *none of the above* and *fill-in-the-blank* might initially appear to be easy to achieve near perfect accuracy on, they can give both the rule-based and GPT-4 methods difficulty. For instance, the rule-based method, which uses keyword matching, might not properly detect *none of the above* if there is a spelling mistake or if there are extra words amongst the given option. Similarly, GPT-4 was often overzealous at detecting these flaws, as at times it interpreted different answer options as effectively containing text akin to *none of the above*, despite it not explicitly being an option.

9.5 Limitations & Future Work

We identified several limitations to our study that may be addressed in future research. First, our study relies on several datasets of student-generated questions, whose quality may vary by the subject area and individual students. Analyzing educational MCQs from other domains that contain a different variety of flaws could lead to more holistic and generalizable findings. It should be noted that the classification of questions in our study is inherently subjective due to the nature of human evaluation. To mitigate this, we employed a standardized IWF rubric and achieved a high inter-rater reliability (IRR) for each criteria. However, it is possible that different evaluators may arrive at different results. The code implementation used to identify the item-writing flaws could be adjusted to achieve different results. For example, variations in threshold for cosine similarity, utilizing an alternative implementation of a method from a different library, or rewording the GPT-4 prompts could affect the outcome.

Finally, the use of GPT-4 poses challenges with replicability, despite providing the prompts and default hyperparameters used in this research. One challenge is that the output of GPT-4 still requires human evaluation to interpret and verify what the model intended, as even when it is prompted for specific phrasing, it may still respond in a conflating manner. Another related challenge is that the model is both inherently random to some degree and still under development, meaning at a future point in time it might perform differently given the same tasks as this research. In order to promote transparency and reproducibility of our research, we have open-sourced our code. This allows for full visibility into the logic used for classifying each item-writing flaw and maintains interpretability so that other researchers can easily make any desired modifications. A promising future direction of this work is to both improve the classification accuracy of these flaws and extend the automatic methods to also provide suggestions for addressing the identified flaws.While GPT-4 may have not been as accurate as the rule-based method for identifying the flaws, it can provide explanations and suggest improvements to the questions.

9.6 Conclusion

In this paper, we proposed a novel rule-based method for automatically evaluating the quality of educational multiple-choice questions using criteria from the Item-Writing Flaws rubric. The results indicate that the rule-based method accurately assesses the quality of student-generated questions across multiple distinct subject areas and highlights the occurrence of different flaws in questions across these domains. It outperforms GPT-4 in applying the Item-Writing Flaws rubric across all four domains when compared to human evaluation. Both automated methods demonstrate how certain flaws may be easier or harder to identify, depending on the subject area. We contribute a categorization and comparison of item-writing flaws found in student-generated questions across four different subject areas. These results provide a valuable baseline performance measure for future research. This work also opens further opportunities for developing open and interpretable methods for evaluating educational questions by pedagogical values.

Chapter 10
Positing a New Method for Educational Question Evaluation

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Eamon Costello, Huy A. Nguyen, and John Stamper. "An Automatic Question Usability Evaluation Toolkit." In *International Conference on Artificial Intelligence in Education*, pp. 31-46. Cham: Springer Nature Switzerland, 2024.

10.1 Introduction

Multiple-choice questions (MCQs) are the most commonly utilized assessment format across educational settings, spanning both traditional classroom environments and digital e-learning platforms [66]. Their versatility allows for assessing a broad spectrum of learning outcomes, ranging from simple recall to complex analytical skills, in many learning domains [152]. Besides offering grading efficiency and objectivity, MCQs enable the targeting of specific misconceptions through carefully crafted alternative answer options, known as distractors. However, the development of high-quality MCQs demands a rigorous approach to ensure reliability, validity, and fairness, essential for accurately measuring learners' knowledge and competencies [207].

Recent advances in natural language processing (NLP) have sought to alleviate the burden and time-consuming nature of MCQ authoring, enabling the rapid generation of questions at scale. These technologies facilitate the generation of hundreds of MCQs within minutes from sources such as document files or direct text requests [141]. Despite these advances, the rise in machine-generated MCQs has not uniformly translated to an improvement in quality. Machine-generated questions produced by state-of-the-art large language models (LLMs) often mirror the inaccuracies commonly found in human generated questions [74]. Such methods raise concerns regarding trust, authenticity, and diversity, potentially leading educators to be hesitant about adopting them without comprehensive evaluation [103].

Among the various MCQ evaluation techniques proposed in the literature, human judgment remains the gold standard, but typically faces challenges with subjectivity, time efficiency, and scalability [170]. Commonly used NLP metrics such as BLEU or METEOR, on the other hand, are much more efficient and scalable, but tend to focus on superficial features like readability and fail to align with human assessments or evaluate the pedagogical value of MCQs [125]. The effectiveness of MCQs are only as good as their design, requiring rigorous evaluation to ensure they serve as effective tools for assessing learning.

To address this gap, our research aims to establish a standardized and rigorous automated technique for MCQ evaluation. We begin by demonstrating the limitations of current NLP-based evaluation metrics, highlighting their lack of correlation with common errors found in MCQs. Then we introduce an automated evaluation technique, Scalable Automatic Question Usability Evaluation Toolkit (SAQUET), designed for comprehensive and standardized quality assessment of MCQs across multiple domains. Leveraging the 19 criteria of the Item-Writing Flaws (IWF) rubric [221], a proven and standardized instrument, SAQUET evaluates the structural and pedagogical quality of MCQs. We evaluate SAQUET across two datasets encompassing 271 MCQs from five diverse fields: Chemistry, Statistics, Computer Science, Humanities, and Healthcare.

The primary contributions of our work include: (1) providing empirical evidence on the inadequacy of prevalent MCQ quality evaluation metrics; (2) introducing SAQUET, an open-source tool capable of domain agnostic MCQ evaluation; and (3) compiling the most extensive and varied open dataset of MCQs annotated with IWF, providing opportunity for future research in educational assessment.

10.2 Methods

10.2.1 Item-Writing Flaws (IWF) Rubric

In our study, we adopted the 19-criteria IWF rubric, a tool that has been validated and employed in prior research [51, 160, 184, 221]. The rubric is designed to be universally applicable across domains, encompassing both pedagogical considerations and factors related to human test-taking abilities. Unlike traditional metrics that primarily assess readability, the IWF rubric includes criteria that address a broader range of question quality aspects, such as unintentional hints, cues, and modality. Table 10.1 outlines each of the 19 criteria, providing guidance on avoiding specific flaws and ensuring adherence to the rubric's standards. Previous research indicates an MCQ with zero or one IWF can generally be considered acceptable for use, particularly in contexts such as formative assessments [221]. Conversely, an MCQ that exhibits two or more IWFs is classified as unacceptable for use. However, instructors might prioritize avoiding specific IWFs based on their use cases to align best with their learning objectives.

10.2.2 Technical Overview of SAQUET

Previous efforts to automate the application of the IWF rubric have explored two main strategies, using either a rule-based approach or the well-known GPT-4 model [160]. The rule-based approach demonstrated superior performance to the

GPT-4-based method for most criteria across all domains used in the previous study. Building upon these findings, this current work enhances the rule-based methodology by integrating advanced methods and incorporating selective GPT-4 interventions. One of our primary objectives was not only to improve the quality of criteria classifications, but also to preserve the tool's ability to be applied across various domains, ensuring scalability and rapid processing for a large volume of MCQs. The automatic detection of the 19 IWF criteria outlined in Table 10.1 falls into three distinct categories: text-matching techniques, NLP-based information extraction, and enhancements provided by GPT-4.

Item-Writing Flaw	**An Item Is Flawed If…**
Longest Option Correct	The correct option is longer and includes more detailed information than the other distractors, as this clues students to this option
Ambiguous Information	The question text or any of the options are written in an unclear way that includes ambiguous language
Implausible Distractors	Any included distractors are implausible, as good items depend on having effective distractors
True or False	The options are a series of true/false statements
Absolute Terms	It contains he use of absolute terms (e.g. never, always, all) in the question text or options
Complex or K-type	It contains a range of correct responses that ask students to select from a number of possible combinations of the responses
Negatively Worded	The question text is negatively worded, as it is less likely to measure important learning outcomes and can confuse students
Convergence Cues	Convergence cues are present in the options, where there are different combinations of multiple components to the answer
Lost Sequence	The options are not arranged in chronological or numerical order
Unfocused Stem	The stem is not a clear and focused question that can be understood and answered without looking at the options
None of the Above	One of the options is "none of the above", as it only really measures students ability to detect incorrect answers
Word Repeats	The question text and correct response contain words only repeated between the two
More Than One Correct	There is not a single best-answer, as questions should have 1, and only 1, best answer

Logical Cues	It contains clues in the stem and the correct option that can help the test-wise student to identify the correct option
All of the Above	One of the options is “all of the above”, as students can guess correct responses based on partial information
Fill in the Blank	The question text omits words in the middle of the stem that students must insert from the options provided
Vague Terms	It uses vague terms (e.g. frequently, occasionally) in the options, as there is seldom agreement on their actual meaning
Grammatical Cues	All options are not grammatically consistent with the stem, as they should be parallel in style and form
Gratuitous Information	It contains unnecessary information in the stem that is not required to answer the question

Table 10.1: The 19 Item-Writing Flaw rubric criteria used in this study.

The first category includes eight criteria: *None of the Above*, *All of the Above*, *Fill-In-The-Blank*, *True or False*, *Longest Answer Correct*, *Negative Worded*, *Lost Sequence*, and *Vague Terms*. Given the nature of these criteria, foundational programming techniques like string matching are primarily used for identification. However, to enhance accuracy we implemented several modifications, such as adjusting threshold parameters, incorporating checks for various question formats, expanding the list of keywords for matching, and lemmatizing the text to normalize word forms. For example, the *True or False* criteria underwent significant alterations to accommodate Yes/No questions. The *Fill-In-The-Blank* criteria required adjustments to avoid misclassification of Computer Science MCQs, which often use the underscore character. Improvements like refined pattern matching were applied to the *Lost Sequence* criteria, enabling the detection of cases not identified in the initial dataset.

The second category encompasses five criteria: *Implausible Distractors*, *Word Repeats*, *Logical Cues*, *Ambiguous or Unclear*, and *Grammatical Cues*. These criteria are addressed using foundational NLP techniques, including word embeddings, Named Entity Recognition (NER), and Transformer models like RoBERTa [170]. NER plays a pivotal role in analyzing *Word Repeats*, *Logical Cues*, and *Grammatical Cues* by allowing us to identify and compare nouns and verbs used in the MCQ. This approach enhances our ability to detect grammatical consistency, identify repeated words, and recognize synonyms. For tackling *Ambiguous Information* and *Implausible Distractors*, our attempts to incorporate GPT-4 faced challenges, as its outputs were often excessively critical, leading to a high rate of misclassifications. To address this, we instead integrated additional linguistic metrics, such as query well-formedness scores [8], and leveraged updated word embeddings to refine the evaluation.

The final category includes six criteria: *Absolute Terms*, *More Than One Correct*, *Complex or K-Type*, *Gratuitous Information*, *Unfocused Stem*, and *Convergence Cues*. This category utilizes NLP techniques similar to the previous ones, enhanced by the integration of GPT-4 API calls for additional verification. For example, simple word matching was insufficient for the *Absolute Terms* criteria, as the context in which terms like “impossible” are used needs further analysis by GPT-4 to determine their impact on answer validity. Modifications were applied to the *Convergence Cues* and *Complex or K-Type* criteria, incorporating GPT-4 for final verification check to improve accuracy. The criteria *Unfocused Stem* and *Gratuitous Information*, both of which involve lexical richness [122], benefited from GPT-4 interventions, significantly reducing false positives detected in pilot tests by better evaluating question stems for learner comprehension. Finally, the *More Than One Correct* criteria was enhanced to not only attempt at answering questions but also to discern whether a question allows for multiple correct responses or is a select-all-that-apply type. We have open-sourced the code and datasets used in this work[3].

10.2.3 Datasets

We utilized two datasets of MCQs previously tagged with the IWF criteria to evaluate SAQUET. The first dataset, derived from [51], encompasses MCQs in Computer Science, Humanities, and Healthcare, sourced from prominent MOOC platforms, such as Coursera and edX. The second dataset, from [160], contains student-generated MCQs from Chemistry and Statistics courses. Both datasets contained MCQs with two to five answer choices each. Additionally, both datasets were evaluated by two human experts, with past studies reporting high inter-rater reliability via Kappa scores. Due to IRB permissions and formatting challenges, not all questions from these initial datasets were included in our present study. Additionally, we made minor corrections to address errors in the datasets, such as mislabeled criteria. For example, one adjustment involved reevaluating Computer Science, Humanities, and Healthcare questions to ensure True/False questions were not mistakenly flagged under the Longest Option Correct criteria, particularly when “False” was the correct answer.

For developing SAQUET, we initially used a subset of 25 questions, 5 from each domain, which were not included in the final evaluation dataset. Our final dataset comprised 271 MCQs across the five domains, all tagged with the 19 IWF criteria, offering a varied pool of questions for analysis. This contrasts with previous IWF research, which often focuses on a single domain [184, 221].

[3] https://github.com/StevenJamesMoore/AIED24

10.2.4 Evaluation

To evaluate the effectiveness of commonly employed automated techniques for assessing question quality, we applied five popular linguistic quality metrics to the 271 MCQs in our dataset: perplexity, diversity, grammatical error, complexity, and answerability. Perplexity scores were generated using a GPT-2 language model, aligning with methodologies from recent research [230]. We measured diversity through the Distinct-3 score, which quantifies the average number of unique 3-g per MCQ [129]. Grammatical errors were identified using the widely recognized Python Language Tool [169], tallying the grammatical inaccuracies in each question as done in prior research [189]. For complexity assessment, we adopted Bloom's Revised Taxonomy, assigning each MCQ a level from 0 (lowest, 'remember') to 5 (highest, 'create'), which serves as a common indicator of complexity and difficulty [122, 152]. A highly precise classifier was employed to automatically determine the Bloom's level for each question [66]. Answerability was evaluated using GPT-4, employing the strategy of the Prompting-based Metric on ANswerability (PMAN) approach [228]. This involved following the strategy of crafting specific prompts that instructed GPT-4 to choose an answer for each MCQ.

For the evaluation of SAQUET, we referenced gold standard human evaluations for our dataset. The overall match rate between our method and the human evaluations is calculated to reflect the general accuracy of our tool in classifying MCQs according to the IWF criteria. To tackle this multi-label classification challenge, we use the exact match ratio, necessitating correct identification of all labels for a match, and Hamming Loss, which calculates the average proportion of incorrect labels, offering detailed insights into our classification's accuracy on a holistic level [80]. We further assess performance using the F1 score of each criteria, which balances precision (the accuracy of positive predictions) and recall (the completeness of positive predictions) . A high F1 score indicates both high precision and high recall, signifying effective identification of an IWF without excessive false positives or negatives. The micro-averaged F1 score aggregates outcomes across all criteria, offering a consolidated view of performance for the entire dataset [132]. Analysis is conducted not just on the aggregate dataset, but also segmented by domain. This allows us to identify domain-specific performance variations and areas for refinement. Where possible, we compare our results with metrics reported in prior studies using similar datasets and evaluation metrics, providing context for SAQUET's performance [51, 160].

10.3 Results

10.3.1 Limitations of Traditional Metrics in Evaluating Educational MCQs

For each of the five domains, we categorized the MCQs into two groups: one group includes MCQs with zero or one IWF and the other comprises MCQs with two or more. This classification helps differentiate between questions that are considered acceptable (zero or one IWF) and those deemed unacceptable (two or more IWF), thereby allowing for a more precise analysis given the constraints of our dataset in accordance with previous research [160, 221]. We then assessed these questions using five linguistic quality evaluation metrics, as detailed in Table 10.2. Our analysis revealed that, across all metrics, the performance of MCQs in each domain either matched or exceeded ones found in recent research. For comparison, [35] reported that human generated MCQs, based on Wikipedia articles and science textbooks, had average perplexity scores of 18 to 84 and diversity scores between .78 and .82. Similarly, [189] determined that the average answerability score for human generated MCQs, on the topic of middle and high school reading comprehension, was .726.

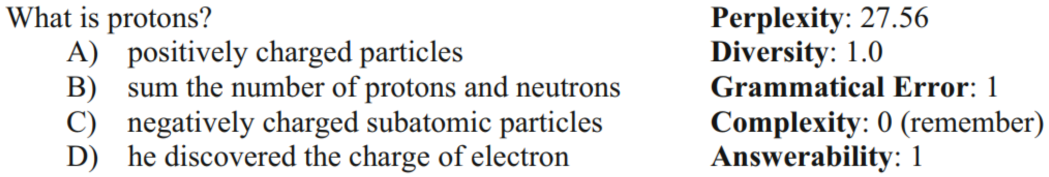
Domain	IWF	N	Perplexity ↓	Diversity ↑	Grammatical Error ↓	Cognitive Complexity ↑	Answerability ↑
Chemistry	0-1	35	47.65	0.961	0.400	0.057	0.743
	2+	15	57.46	0.962(^)	0.333(^)	0.133(^)	0.733
Statistics	0-1	32	46.02	0.928	0.375	0.719	0.531
	2+	18	27.51(^)	0.888	0.444	1.333(^)	0.611(^)
Computer Science	0-1	62	30.73	0.927	2.129	1.145	0.806
	2+	38	41.56	0.917	3.605	1.500(^)	0.605
Humanities	0-1	18	47.64	0.955	0.375	1.313	0.875
	2+	6	28.24(^)	0.939	0.375	1.250	1.000(^)
Healthcare	0-1	25	30.25	0.955	0.400	1.200	0.960
	2+	22	27.72(^)	0.957(^)	0.182(^)	1.682(^)	0.909

Table 10.2: Comparison of five common evaluation metrics for question quality across five domains, categorized by IWF Count. A circumflex (^) denotes a superior score achieved by questions with a higher IWF count in each metric.

Our analysis revealed that student-generated questions in the Chemistry and Statistics domains had relatively high perplexity scores, but in Statistics, questions with 2+ IWFs exhibited a lower perplexity. The diversity metric revealed a ceiling effect, where variations are minimal across different question sets from all domains. High diversity scores are expected, as the MCQs were sourced from diverse origins and authors, such as MOOCs or digital textbooks. The impact of IWFs on a question's answerability varied, where in some cases the presence of IWFs did not reduce, and might have even enhanced, the likelihood of the LLM to correctly answer the questions.

Grammatical errors were relatively low across all domains except Computer Science, where the code syntax posed unique challenges for this criteria, contributing to higher error rates [66]. Interestingly, in both Chemistry and Healthcare, questions with more IWFs (2+) showed a lower average number of grammatical errors, suggesting a nuanced relationship between IWF count and grammatical precision. Initially we expected questions with fewer IWF would have fewer grammatical mistakes, but those may have been overlooked by the human evaluators. Cognitive complexity, measured by Bloom's Revised Taxonomy levels, was also generally higher for questions with 2+ IWFs across all domains except for Humanities, where the difference was marginal, indicating these questions with more flaws tend to engage higher-order cognitive skills.

These findings demonstrate the potential for commonly used metrics to paint an overly optimistic picture of question quality. Even questions with multiple flaws can score well on perplexity, diversity, and grammatical precision, suggesting they are well crafted and clear. However, this can be misleading, as these metrics may not capture deeper issues such as false information, incorrect assumptions, or inaccuracies in content. For example, Figure 10.1 shows a question that achieved an acceptable evaluation across all five metrics, yet it is clearly a poorly student generated question that contains three IWFs: *implausible distractors*, *logical cues*, and *grammatical cues*.



What is protons?
A) positively charged particles
B) sum the number of protons and neutrons
C) negatively charged subatomic particles
D) he discovered the charge of electron

Perplexity: 27.56
Diversity: 1.0
Grammatical Error: 1
Complexity: 0 (remember)
Answerability: 1

Figure 10.1: A student generated MCQ from the Chemistry dataset consisting of three IWFs on the left, with the associated linguistic quality evaluation metrics on the right.

10.3.2 Performance of Automated IWF Classification Across Domains

The 19 IWF criteria were automatically applied to all 271 MCQs for a total of 5,149 classifications. While the overall accuracy is slightly skewed due to most of the questions containing a few flaws and thus being classified as 0 for a given criteria, the total accuracy was 94.13%, which treats each criteria classification individually. We achieved an exact match ratio of 38%, which indicates that 103 of the questions were evaluated the same across all 19 criteria between SAQUET and the different human evaluators. The Hamming Loss was 5.9%, indicating a small amount of misclassification regarding the flaws. While we only used half of the data from [160] consisting of 100 MCQs, it is our closest comparable. As such, compared to their leading rule-based method, we achieved a 3.26% overall classification accuracy improvement, a 13% higher exact match ratio, and 3.1% lower Hamming Loss.

On average, SAQUET (*M=1.75, SD=1.26*) was more likely to classify a MCQ as having more IWFs compared to the human evaluators *(M=1.31, SD=1.11*). The most IWFs assigned to a single question by both was 5. In Table 10.3, we present the IWF classifications from the human evaluators compared to SAQUET for all five domains.

Item-Writing Flaws		Chemistry (50)		Statistics (50)		Computer Science (100)		Humanities (24)		Healthcare (47)	
		Hum	SAQ	Hum	SAQ	Hum	SAQ	Hum	SAQ	Hum	SAQ
Longest Option Correct	N	5	8	3	7	27	27	8	8	16	15
	F1	0.77		0.60		0.96		1.00		0.97	
Ambiguous Information	N	12	12	14	18	12	21	0	2	2	0
	F1	0.58		0.50		0.24		0.00		0.00	
Implausible Distractors	N	9	8	8	6	3	15	3	7	8	3
	F1	0.24		0.86		0.33		0.20		0.00	
True or False	N	2	2	1	0	9	10	4	4	11	11
	F1	1.00		0.00		0.95		1.00		1.00	
Absolute Terms	N	2	1	0	1	9	6	9	9	5	4
	F1	0.67		0.00		0.40		0.89		0.44	
Complex or K-type	N	2	4	4	8	15	12	0	1	4	5
	F1	0.67		0.67		0.81		0.00		0.89	
Negatively Worded	N	0	0	2	4	10	14	0	1	11	11
	F1	-		0.67		0.83		0.00		0.91	
Convergence Cues	N	2	3	9	7	7	11	0	0	1	4
	F1	0.00		0.63		0.44		-		0.00	
Lost Sequence	N	3	3	14	15	2	2	0	0	0	0
	F1	1.00		0.97		0.50		-		-	
Unfocused Stem	N	0	1	8	10	8	5	0	0	0	0
	F1	0.00		0.89		0.62		-		-	
None of the Above	N	6	5	1	1	6	6	0	0	0	0
	F1	0.91		1.00		1.00		-		-	
Word Repeats	N	1	1	1	1	7	11	0	0	4	11
	F1	1.00		1.00		0.56		-		0.53	
More Than One Correct	N	0	2	0	11	8	24	3	10	1	17
	F1	0.00		0.00		0.38		0.46		0.11	
Logical Cues	N	4	3	2	1	2	8	0	0	0	1
	F1	0.29		0.67		0.00		-		0.00	
All of the Above	N	1	1	1	1	2	2	0	0	2	3
	F1	1.00		1.00		1.00		-		0.80	
Fill in the Blank	N	2	2	0	0	2	2	0	0	2	2
	F1	1.00		-		1.00		-		1.00	
Vague Terms	N	0	0	0	1	3	2	0	1	3	3
	F1	-		0.00		0.80		0.00		1.00	
Grammatical Cues	N	2	1	3	1	0	1	0	1	0	1
	F1	0.67		0.00		0.00		0.00		0.00	
Gratuitous Information	N	0	2	3	5	0	3	0	2	0	0
	F1	0.00		0.50		0.00		0.00		-	
Micro-Averaged F1		0.59		0.65		0.62		0.66		0.67	
IWF totals		53	59	74	98	132	182	27	46	70	91

Table 10.3: The number of identified flaws (N) and F1 performance scores for human evaluations (Hum) versus SAQUET (SAQ) across the five domains. A dash

(-) signifies the absence of a flaw in a domain as determined by human evaluation, precluding F1 score calculation.

The F1 scores reveal the effectiveness of SAQUET across the five domains for each criterion. Compared to the rule-based implementation in [160], our approach improved the F1 score across multiple criteria for Chemistry and Statistics questions. Performance on the *None of the Above* criteria was notably strong, as reflected by high F1 scores, indicating precise classification with minimal misclassifications. Other criteria, such as *More Than One Correct*, showed subpar performance across all domains, with frequent incorrect classifications and often overestimating its presence. The micro-averaged F1 scores provide a consolidated view of SAQUET's accuracy across all 19 criteria and allow for a domain-wise comparison of classification efficacy.

Taking the categorization of all MCQs as acceptable (zero or one IWF) or unacceptable (two or more IWF), we compared the SAQUET's classifications with those made by human evaluators. This comparison aimed to see if the overall categorization matched, despite potential misclassifications of specific IWF criteria. Figure 10.2 presents a confusion matrix for this acceptability classification, indicating human evaluators deemed 168 questions acceptable and 103 questions unacceptable. SAQUET matched 204 of these MCQ categorizations, with 112 classified as acceptable and 92 as unacceptable, achieving a 75.3% match rate with human evaluations.

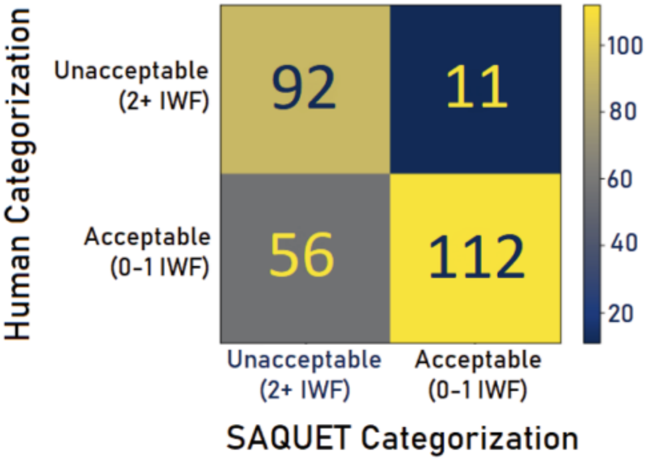


Figure 10.2: A confusion matrix for the categorization of questions as acceptable or unacceptable based on their IWF by the human evaluation and SAQUET.

10.4 Discussion

Our results demonstrate that traditional metrics used for assessing the quality of questions, especially multiple-choice, might not adequately reflect their true quality. We observed that questions with various errors, indicated by Item-Writing

Flaws, which could either simplify the answering process for students or lead to confusion, often receive high scores from commonly used linguistic quality metrics. To address this gap, we introduced SAQUET, a method designed to capture these more complex aspects of question quality while remaining automated and scalable. By benchmarking against human expert evaluations, we show that SAQUET has the potential to provide a more precise and detailed assessment of question quality compared to these linguistic quality metrics. Furthermore, our contribution to the field of assessment quality evaluation research extends to making both SAQUET and our comprehensive dataset publicly available[3].

Recent efforts in NLP have aimed to shift away from traditional readability metrics like BLEU, METEOR, or ROGUE when evaluating the quality of MCQs, yet these metrics continue to be employed in recent works [25, 35, 170]. In our study, we explored alternative linguistic quality metrics (perplexity, diversity, grammar, complexity, answerability) that are also commonly used and offer a different approach to question evaluation, particularly in response to the inadequacies of previous readability metrics [125, 144, 228]. Our findings reveal that even questions with obvious flaws can be evaluated as higher quality according to these metrics. This discrepancy may still hold for machine generated questions from older models, but the improved linguistic capabilities of recent LLMs mean that more machine generated questions are likely to be deemed high quality by these standards. Recent studies have pointed out that despite the grammatical correctness of LLM outputs, the MCQs generated can suffer from issues like implausible distractors or vague wording [74, 170].

SAQUET has the advantage of operating without training data, addressing the significant challenge of sourcing IWF-tagged question datasets. Although research utilizing the IWF rubric is widespread, access to such datasets is often restricted. Importantly, SAQUET's application extends beyond assessing newly crafted questions; it is equally effective in evaluating existing question sets and machine- or human generated questions alike. This capability allows educators to pinpoint and address flaws in current questions they might be using, potentially adjusting or replacing them to suit their needs. In this study, we achieved an exact match ratio of 38% in a complex multi-label classification task with 19 binary labels, which serves as a strong baseline for future research and evaluation. When compared to human evaluations, SAQUET showed a propensity to identify IWFs more frequently. We prefer this stricter approach of identifying MCQ flaws while prioritizing false positives over false negatives, thereby ensuring only the highest quality questions are utilized for educational purposes.

For the criteria based primarily on text matching, such as *True or False*, *All of the Above*, or *Longest Option Correct*, one might intuitively expect perfect accuracy. However, our findings indicate that these criteria can manifest in nuanced forms, demonstrating the importance of datasets that capture a broader spectrum of these errors. For instance, True/False MCQs might also appear as Yes/No choices or contain explanation text that follows the option, complicating their identification. Similarly, interpretations of what constitutions *Longest Option Correct* can vary among human evaluators, as it did in our study. In Chemistry and Statistics this flaw was applied to questions if the second-longest option was not nearly as long (at least 80%) as the longest. In contrast, for Computer Science, Humanities, and Healthcare, a stricter interpretation was applied that flagged any question where the correct answer exceeded others in length by even a single character.

Other flaws like *More than one Correct*, which relied heavily on GPT-4, presented significant challenges, notably impacting the overall exact match ratio. This flaw saw a misclassification for 50 out of 271 questions (18.5%), making it the most problematic. The challenge arose from GPT-4's difficulty in reliably identifying the correct answer for an MCQ, frequently failing to determine if a single correct option exists. However, this limitation is not inherently negative, as it does not imply the question is flawed, just that the LLM has the inability to solve it [153, 228]. This highlights the ongoing challenge of accurately evaluating complex question criteria and the limitations of current AI in navigating such nuances, further emphasizing the need for refined and open approaches along with diverse datasets in the evaluation process.

10.5 Limitations and Future Work

In our study, we introduced SAQUET, an automated method for evaluating questions, employing multiple criteria that leverage LLMs like GPT-4. While outperforming traditional automatedMCQ evaluation metrics, this approach comes with inherent limitations, including the black box nature of LLMs, their potential for unanticipated changes, and the risk of bias in their outputs. To mitigate these issues and enhance this work's reliability and cost-effectiveness, we utilized a specific version of GPT-4 through the `gpt4-0125-preview`[4] API. This approach aimed to standardize the evaluation process and ensure reproducibility by generating consistent outputs from predefined prompts.We further supported transparency and reproducibility by open-sourcing our code1. Expanding our dataset to include a greater number and diversity of questions across additional domains would likely reveal further limitations and areas for improvement in our current evaluation criteria.

For future work, we aim to enhance the evaluation techniques for the 19 IWF criteria, with a particular focus on those that currently show weaker performance. Acquiring additional datasets of MCQs annotated with IWFs will be crucial in validating and demonstrating the effectiveness of our method. We encourage educators, researchers, and practitioners to engage with our work, offering their

[4] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

insights and improvements to refine the criteria further, as we have done. Such collaboration would contribute to developing a more educationally robust metric enriched by collective expertise. As LLMs advance, we anticipate that our methodology will too, achieving greater accuracy for certain criteria and providing detailed feedback on how to correct identified flaws.

10.6 Conclusion

In this study, we highlight the limitations of current metrics for assessing question quality, particularly their oversight of deeper question attributes beyond mere surface characteristics. Through analyzing a dataset of MCQs spanning five varied domains, we illustrate that these prevalent linguistic quality metrics fall short in effectively differentiating between flawed and flawless questions. This gap demonstrates the need for a novel metric capable of comprehensive question quality evaluation. In response, we refined an alternative evaluation method that retains both automation and scalability by assessing MCQs against a detailed 19-criteria Item-Writing Flaws rubric. Upon validating this method to our dataset, we demonstrated its effectiveness across various domains and identified the criteria that were most and least effective. Our findings reveal the potential to significantly enhance question quality assessment, paving the way for more accurate and educationally valuable evaluations.

Chapter 11
Crowdsourced, Expert, and AI-Driven Rubric Applications

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Norman Bier, and John Stamper. "Assessing Educational Quality: Comparative Analysis of Crowdsourced, Expert, and AI-Driven Rubric Applications." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 12, no. 1, *forthcoming*. 2024.

11.1 Introduction

Multiple-choice questions (MCQs) and short answer questions (SAQs) are widely utilized in educational assessments due to their versatility across various learning environments [36, 140]. Despite their popularity, creating high-quality and reliable educational assessments is challenging, often requiring significant time and domain specific expertise [47, 86]. Existing tools and methods for crafting and evaluating these questions are not without their issues, capable of producing questions with inherent flaws that are potentially detrimental to their pedagogical value. These flaws can persist in widely used question datasets and across online courses, hindering the student learning process [50, 196]. The gold standard for identifying and correcting these issues traditionally involves expert human judgment [102]. Automated evaluation methods, although less subjective, typically depend on extensive student performance data or focus on superficial metrics like readability, which do not fully capture the educational effectiveness of the questions or correlate with human judgment [18].

Despite the recognized need for human expertise in evaluating the quality of educational assessments, relying solely on such input limits the scalability and efficiency of the process. Crowdsourcing and learnersourcing offer potential solutions by leveraging collective human intelligence on a larger scale, though these methods often involve participants with less expertise [212, 220]. Moreover, recent developments in Large Language Models (LLMs) suggest that AI could mimic humanlike judgment for certain educational tasks, offering a scalable approach for assessing question quality [3, 143].

In response to these challenges, this study compares the effectiveness of multiple crowdsourcing strategies with LLM-based methods for evaluating the quality of 30 MCQs and SAQs across six domains. These evaluations employ two standardized and validated rubrics, examining the assessments' pedagogical validity. We conducted two distinct crowdsourcing tasks, one for MCQs and another for SAQs, to see how well novice contributors could apply these rubrics. Concurrently, we utilized three state-of-the-art LLMs to automate the same evaluation process. By analyzing the wisdom of the crowds [120] this research assesses how closely the majority responses from crowdsourced evaluations align with those generated by LLMs and verified by subject matter experts. This study investigates two primary research questions: How do the effectiveness and accuracy of rubric applications by crowdworkers, experts, and AI models compare when assessing educational content (RQ1)? How consistent and reliable are quality assessments of MCQs and SAQs within crowdsourced and LLM methods (RQ2)?

Through the investigation of these research questions, this work makes the following contributions: First, it demonstrates the comparative effectiveness and accuracy of rubric applications by crowdsourced workers, experts, and LLMs in evaluating the quality of educational assessments. Second, it provides a detailed analysis of the consistency and reliability of quality evaluations for MCQs and SAQs, highlighting critical trade-offs. Finally, it provides insights into the integration of LLMs in the educational quality evaluation process, proposing a potential hybrid approach that leverages both human expertise and AI to enhance the quality and reliability of educational assessments.

11.2 Methods

To explore the effectiveness and trade-offs between crowdsourced and programmatic LLM-based methods in assessing the quality of educational questions, we conducted a comparative study across two types of questions, multiple-choice questions (MCQs) and short answer questions (SAQs), spanning six subject areas. Our study comprised two experiments: the first evaluated the quality of MCQs using the 19-criteria IWF rubric, applied by various crowdworkers and multiple LLM-based programmatic methods; the second experiment involved a similar evaluation of SAQs using a 9-item rubric. In total, 30 questions were evaluated, 15 MCQs and 15 SAQs, from distinct domains within mathematics, science, and the humanities. All the questions were purely text-based, with no accompanying images or formulas. The 15 MCQs used in this research were sourced from a previous study, where the IWF rubric had already been applied [51]. These MCQs were extracted from introductory online courses in Philosophy, Statistics, and Chemistry.

The five SAQs related to Chemistry were obtained from a separate study involving an online introductory Chemistry course [158]. We selected five SAQs each from online Calculus and Team Collaboration courses at a university on the U.S. East Coast. The Team Collaboration course covers communication, teamwork, and conflict management. These questions were selected by two domain experts, who identified potential flaws within them. For all 15 of the SAQs, the experts applied the 9-item SAQ rubric to evaluate these questions. To assess the consistency of their evaluations, we calculated the inter-rater reliability using Cohen's Kappa [148]. The overall Cohen's Kappa score was 0.79, indicating substantial agreement between the raters across the entire rubric. Further details about these questions are available in Table 11.1.

Domain	Type	Number	Number of Flaws
Philosophy	MCQ	5	10
Statistics	MCQ	5	11
Chemistry	MCQ	5	10
Team Collaboration	SAQ	5	12
Calculus	SAQ	5	6
Chemistry	SAQ	5	11

Table 11.1: Information about the 30 questions used in this research.

11.2.1 Item-Writing Flaws Rubric

To evaluate the MCQs, we engaged both crowdworkers and LLMs to apply the IWF rubric. The IWF rubric encompasses 19 criteria specifically designed to assess the quality of educational MCQs. This version of the rubric has been extensively used and validated in previous research, particularly within STEM fields [31, 196]. The criteria cover various aspects of the questions, including the question text, answer choices, and the correct option, ensuring a comprehensive evaluation of each component. While expertise is not required to apply this rubric, certain criteria, such as identifying implausible distractors or logical cues, may benefit from domain knowledge as well as an understanding of assessment creation. A detailed list of the 19 IWFs and their definitions is provided in Table 11.2.

Item-Writing Flaw	Definition
Absolute Terms	Use of definitive words like "always" or "never" that can make a statement unequivocally true or false.
All of the Above	Inclusion of an option that suggests selecting all previous options, often giving away the correct answer.
Ambiguous Information	Unclear or vague content that can lead to multiple interpretations.
Convergence Cues	Clues within the question or options that guide test-takers to the correct answer.

Logical Cues	Answer choices that can be deduced logically rather than through knowledge of the subject.
Complex or K-type	Use of complex formats like multiple true-false questions within a single item, which can confuse test-takers.
Fill in the Blank	Questions requiring the test-taker to provide a missing word or phrase, which can be too open-ended.
Grammatical Cues	Grammatical inconsistencies between the stem and the correct answer that can hint at the correct choice.
Gratuitous Information	Unnecessary details that do not contribute to the question, potentially distracting the test-taker.
Implausible Distractors	Option choices that are obviously incorrect, making the question too easy.
Longest Correct	The correct answer is noticeably longer than the distractors.
Lost Sequence	Options that are not presented in a logical or sequential order, causing confusion.
More than One Correct	Multiple correct answers when only one is expected, causing ambiguity.
Negative Wording	Use of negative phrases like "Which of the following is NOT..." that can confuse test-takers.
None of the Above	Including an option that invalidates all other choices, which can be misleading.
True or False	Avoid simplistic questions using true or false, as they reduce the depth of assessment.
Unfocused Stem	The question stem is not clear or concise, leading to confusion about what is being asked.
Vague Terms	Use of unclear or imprecise terms that can be interpreted in multiple ways.
Word Repeats	Repetition of words or phrases in the stem and the correct answer, providing unintended hints.

Table 11.2: Definitions for each of the 19 criteria used to identify common item writing flaws in educational MCQs.

11.2.2 Short Answer Question Rubric

To evaluate the SAQs, we employed both crowdworkers and LLMs to apply a 9-item rubric. The rubric was from two previous studies that used a version of it for evaluating STEM questions [91, 217]. We adjusted the rubric by combining elements from both studies to minimize the inclusion of overly subjective criteria. However, unlike the more objective IWF rubric, this SAQ rubric still contains criteria that can be interpreted differently. The final version of the SAQ rubric used in this study is detailed in Table 11.3. It lists the criteria labels along with corresponding yes-or-no questions that assess whether each criterion is met or violated. It is important to note that the answer to the SAQ or any other associated metadata is not required for applying this rubric's criteria during the evaluation process. Like the IWF rubric, applying this rubric may be easier for evaluators with domain knowledge, particularly for criteria that specify [*specific domain*].

Item-Writing Flaw	**Definition**
Understandable	If you were a student in a [specific domain] course, could you clearly understand this question without additional explanations?
Domain Related	Is the question related to [specific domain]?
Grammatical	Is the question grammatically correct and free of language errors?
Focus	Is the question specific and focused on a single concept or topic?
Conciseness	Is the question concise and free of unnecessary information?
Fairness	Is the question culturally neutral and free from any biases that might disadvantage any group of students?
Cognitive Level	Does the question require students to apply higher-order thinking skills rather than simply recalling facts?
Central	Is being able to answer the question important for understanding the topics covered by a course in [specific domain]?
Would You Use It	If you were a teacher working with content related to this question in your course, would you include this question in the course?

Table 11.3: Definitions for each of the 19 criteria used to identify common item writing flaws in educational MCQs.

11.2.3 Participants

We recruited participants using two popular crowdsourcing platforms, Amazon's Mechanical Turk (MTurk) and Prolific [67]. On each platform, participants received identical instructions for the task, which involved completing a survey. In this survey, they evaluated five questions at a time, applying the appropriate rubric based on the type of question being assessed.

To evaluate the 15 MCQs using the 19-criteria IWF rubric, we adapted it into yes-or-no questions for crowdworkers to assess whether each MCQ violated specific criteria. Participants were presented with each MCQ, which included the question text and four answer choices. They were informed that the first choice, option A, was the correct answer. This process is conducted on five distinct MCQs drawn from introductory courses in Philosophy, Statistics, or Chemistry. The crowdworkers evaluate each of the 19 criteria for one MCQ before moving on to the next, completing evaluations for a total of five MCQs.

Similar to the IWF task for MCQs, the evaluation process for SAQs involved applying the 9-criteria SAQ rubric to each question, structured as a series of nine yes-or-no questions. Crowdworkers assessed each SAQ individually, completing evaluations for five SAQs sequentially. Each set of SAQs covered one of the three domains used in this experiment: Team Collaboration, Calculus, or Chemistry.

We recruited two distinct groups of participants: novice crowdworkers from MTurk and those with some domain expertise from Prolific. All participants were over 18 years old, self-reported as expert or native English speakers, and were compensated at a rate of at least $18 per hour for their time. The tasks were designed to be efficient: the average completion time for the MCQ task was 14 minutes and 36 seconds, while the SAQ task took an average of 9 minutes and 2 seconds. Upon completing the task, participants were asked to self-report their understanding using a five-point Likert scale and to provide any written feedback. All participants from both platforms reported scores of 4 or 5, indicating a high level of full understanding of the task.

Amazon's Mechanical Turk We utilized MTurk to recruit 11 unique crowdworkers for each of the six question evaluation tasks, totaling 66 participants. To ensure high-quality contributions without severely limiting our participant pool, we established qualifications requiring that each crowdworker had an overall approval rate greater than 95% before they could participate in our study. These participants were considered novice, as none reported having professional experience in the domains of the questions or in education more broadly.

Prolific We recruited 18 unique crowdworkers from Prolific, assigning three crowdworkers to each of the six tasks. Each participant possessed at least a bachelor's degree in the domain relevant to their assigned questions. For example, the crowdworkers evaluating the five Calculus SAQs held degrees in mathematics. This group was considered more advanced and knowledgeable than those from MTurk, owing to their specialized educational backgrounds. Previous research has shown that Prolific generally attracts a higher skilled audience capable of delivering superior results [67]. Due to these factors and associated cost considerations, we decided to limit the number of Prolific crowdworkers to three per task, with each group evaluating a set of five questions from one of the six domains.

11.2.4 Application of Large Language Models

We employed three LLMs, GPT-4, Gemini 1.5 Pro (Gemini), and Claude 3 Opus (Claude), to programmatically apply the two rubrics to our question set [104]. These three models were chosen for their strong performance benchmarks, widespread popularity at the time, and easy API access. For the automated application of the IWF rubric, we utilized an established automated method that applies various NLP techniques tailored to each of the 19 criteria [160]. This method has been previously applied in several studies involving MCQs in domains such as Biology and Algebra [14, 154]. While this automated method incorporates the use of an LLM, for our current study, we varied which LLM was employed in each evaluation cycle to assess their relative effectiveness.

For the SAQs, our approach mirrored that of previous studies which have successfully used LLMs to apply rubric criteria to educational content [101, 236]. Adopting the LLM prompting strategy of having it assume the role of an expert, we assigned the LLMs the role of an experienced instructor tasked with evaluating the quality of educational content [136]. Given that the 9- item SAQ rubric consists of yes-or-no questions corresponding to each criterion, we used these questions as prompts for the LLMs, inputting both the rubric question and the text of the SAQ for evaluation.

The total cost and time required to evaluate all 15 MCQs and 15 SAQs using these methods are detailed in Table 11.4. We used a single iteration of LLM prompting for each set of questions, without running multiple iterations or combining outputs, to align with methods used in previous research.

Method	Type	Cost (USD)	Time (seconds)
GPT-4	MCQ	0.21	28
GPT-4	SAQ	0.72	77
Claude 3 Opus	MCQ	0.13	504
Claude 3 Opus	SAQ	0.24	1557
Gemini 1.5 Pro	MCQ	0.04	33
Gemini 1.5 Pro	SAQ	0.12	79

Table 11.4: The cost and time of GPT-4, Gemini 1.5 Pro, and Claude 3 Opus in applying the IWF and SAQ rubrics.

11.2.5 Data Analysis

For each evaluation task, we assessed the crowdworkers' ability to effectively apply the specified criterion from each rubric using a consensus-based approach. Specifically, we adopted the majority response as the representative outcome for each criterion. For example, in the Calculus MCQ task, if six out of eleven crowdworkers indicated that a question violated the first criterion of the IWF rubric, this majority view was taken as the crowd's consensus. This method, often referred to as the “wisdom of the crowd", is a widely used technique for aggregating responses from crowdsourcing platforms [120].

For our comparison of accuracy between the crowdsourced and LLM methods, we referred to the human evaluations within our dataset. We addressed this multi-label classification challenge using the Exact Match ratio, which requires correct identification of all labels for a question to be considered a match, and the Hamming Loss, which calculates the average proportion of incorrect labels, providing detailed insights into our classification's holistic accuracy [185]. Performance was further assessed using the Micro F1 score for each criterion, which combines precision (the accuracy of positive predictions) and recall (the completeness of positive predictions) to deliver a measure of each method's effectiveness in accurately classifying each specific criterion of the rubrics [241]. A high Micro F1 score indicates both high precision and high recall, indicating effective identification of criteria with minimal false positives or negatives.

Additionally, we evaluated the Macro F1 score, which averages the F1 scores computed for each criterion independently, showing how uniformly the method performs across diverse categories without being influenced by the frequency of each criterion [241]. Finally, we utilized the Jaccard Index as another metric [79]. This index measures the intersection over the union of the predicted and actual labels at an aggregate level, offering a direct indicator of the overlap between the two sets. This metric is valuable for assessing the overall effectiveness of the classification in scenarios where accurate positive identifications are essential output integrity.

11.3 Results

11.3.1 Crowdsourcing Outperformed the LLMs

MCQ Quality The 19 IWF criteria were applied to all 15 MCQs, resulting in a total of 285 classifications. The evaluation metrics for our five assessment methods, calculated by comparing them to the ground truth MCQ labels previously provided by two experts, are presented in Table 11.5.

For raw accuracy, MTurk shows the highest exact match ratio at 33%, indicating it has the highest proportion of correct predictions, exactly matching expert labels for five of the fifteen questions. It also has the lowest Hamming

Loss at 7%, indicating a small amount of misclassification regarding the flaws. Due to having the most precise prediction with fewest incorrect labels, it is the best performing method for the IWFs. However, Prolific excels in the other three metrics, demonstrating the best balance of precision and recall across all 19 IWF criteria.

In comparison, the three automated methods perform worse than the crowdsourcing methods of MTurk and Prolific. Among the automated methods, GPT-4 performs the best, achieving moderate performance across all evaluation metrics. The poorest performer overall is Gemini, which has the lowest scores across all five evaluation metrics.

Method	Exact Match	Hamming Loss	Micro F1	Macro F1	Jaccard Index
MTurk	**0.333**	**0.070**	0.643	0.385	0.474
Prolific	0.200	0.081	**0.667**	**0.535**	**0.500**
GPT-4	0.200	0.091	0.567	0.370	0.395
Gemini	0.067	0.140	0.444	0.316	0.286
Claude	0.133	0.105	0.500	0.319	0.333

Table 11.5: Performance of 5 methods at applying the 19-criteria IWF rubric to 15 educational MCQs.

SAQ Quality The 9-item rubric was applied to all 15 SAQs, resulting in a total of 135 classifications. The evaluation metrics for our five assessment methods, calculated by comparing them to the ground truth SAQ labels provided by two experts, are presented in Table 11.6.

Method	Exact Match	Hamming Loss	Micro F1	Macro F1	Jaccard Index
MTurk	0.200	0.193	0.886	0.875	0.795
Prolific	0.200	**0.163**	**0.897**	0.879	**0.814**
GPT-4	**0.267**	0.185	0.892	**0.882**	0.805
Gemini	0.133	0.193	0.876	0.864	0.779
Claude	0.133	0.244	0.841	0.827	0.725

Table 11.6: Performance of 5 methods on applying the 9-criteria rubric for evaluating educational SAQs.

While GPT-4 achieves the highest exact match ratio, successfully classifying the most SAQs accurately across all 9 criteria, it is not the best overall performer. Prolific stands out by achieving high evaluation metrics, particularly in Micro F1 and Jaccard Index, which indicate a strong balance of precision and recall. Additionally, Prolific has the lowest Hamming Loss of all methods, indicating it is the most accurate in labeling. Similar to its performance in the MCQ evaluation, MTurk also performs well, achieving metrics only slightly lower than Prolific.

The remaining two automated methods, Gemini and Claude, performed poorly by comparison. Claude achieved the worst results across all evaluation metrics. This shows that while some automated methods can be effective, there is significant variation in their performance. Overall, Prolific emerges as the most reliable method for evaluating the quality of educational SAQs, combining high precision and recall with the lowest rate of labeling errors.

11.3.2 Evaluating Method Trade-Offs

While much of the evaluation metrics focused on exact match and Hamming Loss, these do not provide a holistic picture. Exact match is strict and can be skewed by a few incorrect predictions, while Hamming Loss offers only a broad error rate. To provide a more comprehensive evaluation, we use the Micro F1 score, which considers both precision and recall, enabling a more accurate and realistic assessment of each method's effectiveness [241].

IWF Performance We present a comparison of Micro F1 scores for the top three performing methods, MTurk, Prolific, and GPT-4, across all 19 criteria from the IWF rubric, as shown in Figure 11.1.

Less subjective criteria, which are simple enough to be addressed through programmatic string matching, such as *absolute terms*, *all of the above*, and *fill in the blank*, perform highly across all three methods. In contrast, more subjective measures that might be influenced by domain knowledge or instructional design preferences posed a challenge for the three methods. For instance, *ambiguous information* is one of the lower-scoring criteria, especially for GPT-4, indicating a difficulty in handling ambiguity in text. Similarly, *implausible distractors* present a challenge for all three methods, although GPT-4 performs the best in this area despite it requiring domain knowledge. Separating the evaluation by criteria further demonstrates that Prolific consistently achieves the highest performance. Notably, there are multiple criteria where all three methods perform at the highest level, achieving a Micro F1 score of 1.

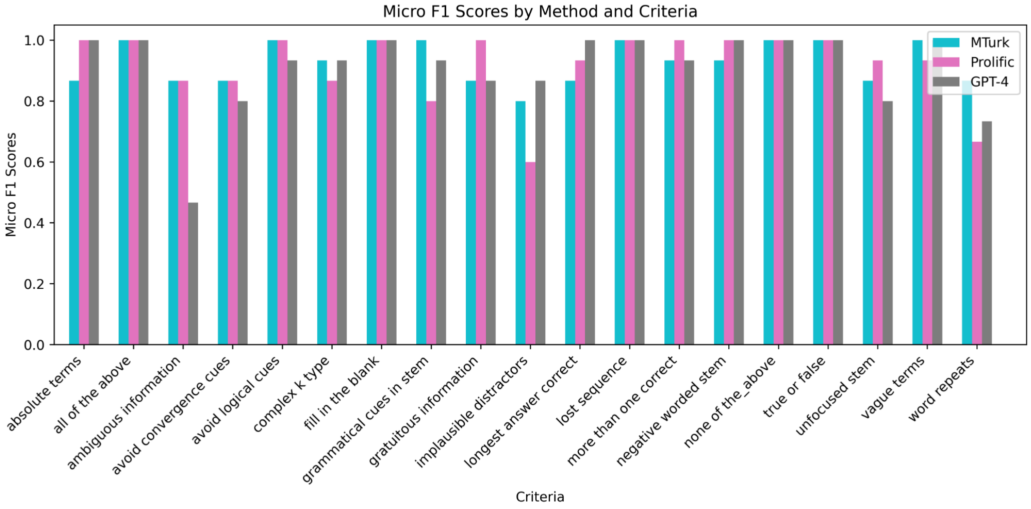


Figure 11.1: Comparison of Micro F1 scores across 19 IWF criteria for MTurk, Prolific, and GPT-4, illustrating the performance of the three top-performing methods in evaluating MCQs.

SAQ Rubric Performance We present a comparison of Micro F1 scores for the top three performing methods, MTurk, Prolific, and GPT-4, across all 9 criteria from the SAQ rubric, as shown in Figure 11.2.

The most subjective criterion, *would you use it*, posed a challenge, as even the experts who reviewed these questions to create our ground truth struggled with this criterion, as it is purely subjective and influenced by many factors. Less subjective criteria, such as the *conciseness* of the text, also had poor performance across all methods.

The *understandable* criteria achieved high performance despite its potential subjectivity and influence from domain knowledge. Compared to the two crowdsourcing methods, GPT-4 achieved superior performance in *cognitive level* and *grammatical* criteria. Machines are typically good at these tasks, as cognitive level can be partly determined by verb usage [17], and grammatical correctness has been a significant focus of LLMs and NLP work [236]. Even though Prolific generally achieved the highest or tied for the highest performance on each criterion, GPT-4 performed quite close to it.

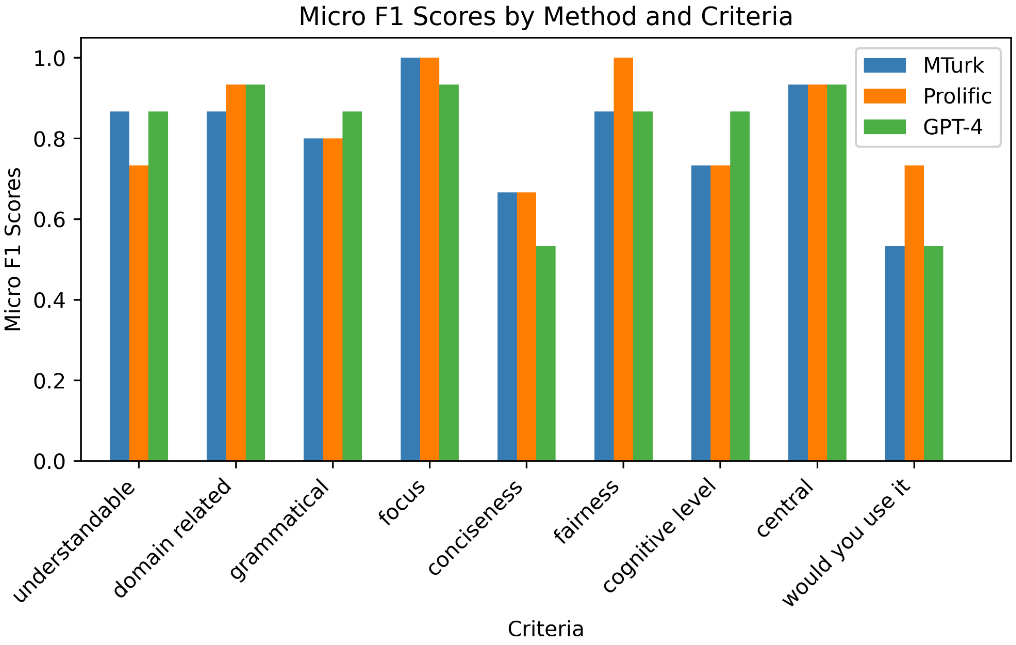


Figure 11.2: Comparison of Micro F1 scores for evaluating SAQs across 9 criteria by the three methods.

11.4 Discussion

In this study, we evaluated 30 questions, five from each of six distinct domains, using two crowdsourcing platforms and three state-of-the-art LLMs to apply two different rubrics. The results indicate that while the human-involved crowdsourcing methods generally outperformed the automated approaches, the LLMs performed comparably well on many criteria and even exceeded human performance on some. Across both types of questions at least one method, automated or crowdsourced, achieved perfect or near-perfect classification for a given criteria in alignment with human expert labels. These findings support the potential for a hybrid approach where human expertise is utilized primarily for the most challenging criteria, while AI handles the more straightforward tasks.

11.4.1 Method Evaluation

We observed that the crowdsourcing methods outperformed the programmatic methods across the two tasks. Specifically, MTurk was the top performer for IWFs and Prolific excelled in the SAQ rubric evaluation. MTurk's workers typically possess less domain knowledge, a factor we controlled by selecting participants with relevant expertise on Prolific [67, 155]. The detailed and less subjective nature of the IWF rubric likely aided MTurk workers by providing sufficient guidance, despite their varied knowledge levels. For the SAQ task, Prolific superior performance is attributable to our targeted recruitment of individuals with relevant academic qualifications. This was crucial since the SAQ task's criteria are inherently more subjective and knowledge-intensive [217].

Regarding automated methods under the same research question, they generally underperformed when compared to the crowdsourced approaches. Notably, the latest iterations of Gemini 1.5 Pro and Claude 3 Opus were less effective than GPT-4, while also taking longer to complete (see Table 11.4). However, across all evaluation metrics, crowdsourced methods consistently outperformed automated ones. Despite the challenges of these 19-item and 9-item multi-label classification tasks, where achieving an exact match required correct labeling of each item, all methods managed to maintain a low hamming loss rate, demonstrating a base level of competency in handling these complex tasks.

11.4.2 Performance Variability

In applying the IWF rubric, both crowdsourcing methods either matched or outperformed the automated GPT-4 process, with the notable exception of the *implausible distractor* criteria. It appears that GPT-4 may have surpassed the crowd in this area due to its ability to quickly identify outliers in data sets [219]. This finding suggests that while human input remains crucial in the question quality evaluation process, automated methods could effectively handle specific criteria where their performance is comparable to that of humans. Implementing such a hybrid approach could reduce the workload for experts or crowdworkers; instead of assessing 19 IWF criteria per question, they might only need to evaluate 5, primarily confirming or refining the outputs from the automated evaluation. This could also lessen the demand for deep domain knowledge, as crowdworkers could focus on verifying the logic behind the AI's classifications, which provides a layer of human oversight to help mitigate the potential bias and errors introduced by the LLM [97].

Furthermore, in the more complex and detailed IWF rubric (19 criteria) compared to the SAQ evaluation (9 criteria), the performance was generally lower. The subjective nature of the *would you use it* criteria posed a particular challenge, especially for programmatic methods. It is difficult for both LLMs and humans to assess such a criterion effectively without substantial contextual information. Even less subjective criteria, like the *conciseness* of SAQs, showed low performance across all methods. This variability could be attributed to the diverse interpretations of *conciseness* among crowdworkers given their unique backgrounds.

11.4.3 Feasibility

Assessing the feasibility of different methods for evaluating educational content, it becomes clear that neither experts nor crowdsourcing are cost-effective options. For instance, while Prolific achieves high results, the time and cost it took to set up the task to evaluate five questions in a typical online course is impractical.

Despite the costs and challenges, automated methods have shown promise, particularly as LLMs continue to advance. Yet, human computation still appears to be the most effective for evaluating MCQs and SAQs. Combining the two by integrating human insights with automated processes could optimize efficiency. For example, the use of GPT-4 could be integrated as part of a hybrid workflow, as it has demonstrated success by achieving perfect Micro F1 scores for several criteria and performing comparably to human evaluators in other aspects. This suggests a combined approach might alleviate some of the burdens on human evaluators by involving them only when necessary.While designing better questions from the start is ideal, there's a practical aspect to consider as well: many existing questions are already available in various banks and online courses [51]. Instead of creating new content from scratch, a more efficient approach could be to evaluate and improve existing questions. Crafting high-quality MCQs and SAQs is a skill that requires time and practice, and even LLMs occasionally produce flawed questions. Recognizing that no method is perfect, leveraging both automated and human resources could enhance the overall quality of educational assessments.

11.4.4 Limitations

The inherent subjectivity associated with human ratings was addressed by employing verified and validated rubrics, yet some level of subjectivity inevitably remains. Additionally, the use of LLMs introduced potential biases related to their training data and algorithms. The task formulation itself, both for the crowdworkers and the LLMs, presented challenges, including the precise wording of rubric criteria and considerations regarding the native language of participants, which could affect their understanding and application of the rubrics. Particularly with LLMs, the various prompt wordings can drastically change the outputs as well, so consistent phrasing and temperature is crucial for reliable results.

11.5 Conclusion

This study explored the effectiveness and reliability of crowdsourced and programmatic methods for evaluating the quality of multiple-choice questions MCQs and SAQs across various educational domains. By leveraging the IWF rubric and a 9-item SAQ rubric, we systematically compared the performance of crowdworkers from MTurk and Prolific with three state-of-the-art LLMs: GPT-4, Gemini 1.5 Pro, and Claude 3 Opus. Our findings reveal that while crowdsourcing can harness wide-reaching human insights, LLMs offer a scalable alternative that approaches the reliability and accuracy of expert judgments. The application of standardized rubrics by both crowdworkers and LLMs highlighted the potential for a hybrid approach, combining the nuanced understanding of human reviewers with the efficiency and consistency of automated systems. This work highlights the trade-offs of each method and demonstrates the feasibility of integrating these approaches to improve the pedagogical value of assessments. As we move forward, refining these hybrid strategies could significantly enhance the way educational content is evaluated, ensuring both the scalability of the evaluation process and the quality of educational assessments.

Chapter 12
Automated Generation and Tagging of Skills to MCQs

> This chapter is based upon the following previously published work:
>
> Moore, Steven, Robin Schmucker, Tom Mitchell, and John Stamper. "Automated Generation and Tagging of Knowledge Components from Multiple-Choice Questions." In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pp. 122-133. 2024

12.1 Introduction

Digital learning platforms, such as MOOCs and interactive online courses, facilitate the mapping of assessments to specific skills or competencies, referred to as Knowledge Components (KCs). These KCs are instrumental in driving learning analytics systems, enabling adaptive content sequencing, and providing precise estimates of student mastery levels [26, 38, 94]. KCs embody the cognitive functions or structures inferred from student performance on related assessments and are more nuanced and fine-grained than broad course learning objectives [116]. The process of associating assessments with KCs, often referred to as skill or concept tagging, generates a comprehensive map of the knowledge conveyed by the platform or course. This mapping, termed Knowledge Component Model (KCM), is crucial for closely monitoring student learning, allowing educators to identify precisely which concepts a student may find challenging based on their assessment performance [191]. Without an accurate KCM, the effectiveness of assessing mastery and implementing adaptive learning strategies may be significantly hindered [179].

While KCMs offer numerous benefits for digital learning platforms, the creation of KCs for each assessment is a time intensive process that demands domain expertise. This usually requires a domain expert, such as the course instructor, to determine the necessary KCs for solving each assessment [45]. This process necessitates the identification of at least one KC per assessment item and can involve identifying up to three or more KCs, depending on the complexity of the assessment, the subject area, and the educational level [28]. Modifying existing KC tags is equally challenging; although evaluating the effectiveness of KCMs is feasible, adjusting it can be as time-consuming as the initial creation process. Evaluation methods such as learning curve analysis and statistical measures such as Akaike Information Criterion (AIC) can provide insights into the effectiveness of the KCMs [118]. However, these methods often require large amounts of data to produce reliable results, and primarily focus on the fit and predictive accuracy of the model, potentially overlooking the contextual and practical applicability of the KCs in diverse educational environments. The continual evaluation and updating of these models remain critical, as many learning platforms in the United States have begun transitioning to common taxonomies, like the Common Core State Standards, necessitating the retagging of content or the alignment of existing mappings with these standards [194]. This re-mapping is not only labor-intensive but also needs to be revisited with each update or change to the common standards, adding to the ongoing workload.

To mitigate the challenges associated with mapping KCs to assessments, a variety of automated solutions employing machine learning (ML) and natural language processing (NLP) have been proposed [78, 180]. These approaches primarily employ classification algorithms, using an existing repository of KCs as reference labels for mapping. However, one critical limitation of these techniques is their reliance on a predefined set of KCs. They are not designed to identify new KCs, but instead depend on a pool of KCs, which may not always be available. The difficulty of automatically generating new KCs lies in ensuring their specificity and relevance to the problem at hand, their relationship with other KCs, and their alignment with the overall course content [226]. Consequently, while the challenge of associating assessments with existing KCs is significant, the task is further complicated by the initial requirement to generate these KCs, a step that previous efforts often overlook. When previous studies involve KC generation, they are frequently unlabeled, necessitating a human to create their descriptive text labels [21, 38].

Recent advancements in large language models (LLMs) have shown promise in automating the generation and tagging of metadata to educational content [200]. To explore this possibility in the context of KCs, we utilize datasets from two different domains: Chemistry and E-Learning, encompassing the higher-education levels of undergraduate and masters. The respective datasets contain multiple-choice questions (MCQs), with each question mapped to a single KC. We leverage GPT-4 [175], a state-of-the-art LLM, to generate a KC for each MCQ by employing two distinct prompting strategies, based solely on the text of the MCQs. We then compare LLM-generated KCs to the original human-assigned ones, which serve as our gold standard labels. For KCs that did not match, we had groups of three human domain experts evaluate the discrepancies to determine their preferred KCs, which often leaned towards the ones generated by the LLM. Additionally, to organize the KCs, we implement a clustering algorithm that leverages the content of the MCQs to group questions that assess the same KCs. Our research presents a methodology that automates the process of generating and tagging KCs to problems across various domains, relying exclusively on the text of the assessments.

The main contributions of this work are: 1) A proposed method for generating KCs for assessments using LLMs, 2) Empirical and human validation of the LLM-generated KCs, and 3) A technique that iteratively induces a KC ontology and that clusters assessments accordingly.

12.2 Methods

12.2.1 Datasets

In this study, we utilized two datasets from higher education MOOCs: one in Chemistry and the other in E-Learning. Each dataset comprises 80 multiple-choice questions (MCQs), with each question offering between two to four answer options. These datasets are structured such that each KC is represented by exactly two MCQs, totaling 40 KCs in each dataset. The ground truth KCs associated with each question were previously identified by domain experts who contributed to developing the content and authoring the courses. A key selection criterion for these questions was that each should be associated with a single KC, ensuring clarity in mapping.

The Chemistry dataset[5] originates from an online course adopted by various universities across the United States as instructional materials for an undergraduate introductory Chemistry course. This content is hosted on a widely used digital learning platform and is often integrated into a flipped-classroom model, supplementing in-person instruction. Similarly, the E-Learning dataset[6] is derived from a master's level course at a university in the eastern United States, utilizing the same digital platform for implementation. Both datasets and the code utilized in this study are available for inspection[7].

12.2.2 Prompting Strategies

To automatically generate KCs, our research employed the `gpt4-0125-preview`[8] API, chosen for its speed, reduced cost, and the consistency it offers. This decision was made to avoid the variability that might arise from continuous updates to the standard GPT-4 API, which could impact the reproducibility of our results. Following recommendations from existing literature [16, 190], we developed two prompting strategies to generate the KCs for each question: the *simulated expert* approach and the *simulated textbook* approach. For both strategies, the only contextual information provided to the language model was the course's educational level (undergraduate or master's) and the

[5] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=4640

[6] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=5843

[7] https://github.com/StevenJamesMoore/LearningAtScale24

[8] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

subject area (Chemistry or E-Learning). For each approach we supplied a MCQ, which included the question text, the correct answer, and the alternative options. This choice was motivated by our aim to develop a method that could be generalized, recognizing that providing extensive contextual information might not always be feasible for certain question banks or assessment content. This is also aligned with previous research that has attempted to generate KCs based solely on the content’s text using automated NLP-based approaches [146, 208]. The exact prompts used for the *simulated expert* strategy are illustrated in Figure 12.1 and the prompts for the *simulated textbook* strategy are in Figure 12.2.

```
“““Simulate three experts collaboratively evaluating a
college level multiple-choice question to determine what
knowledge components and skills it assesses. The three
experts are brilliant, logical, detail-oriented, nit-picky
{subject} teachers. The multiple-choice question is used
as a low-stakes assessment as part of an {context}
{subject} course that covers similar content. Each person
verbosely explains their thought process in real-time,
considering the prior explanations of others and openly
acknowledging mistakes. At each step, whenever possible,
each expert refines and builds upon the thoughts of
others, acknowledging their contributions. They continue
until there is a definitive list of five knowledge and
skills required to solve the question, keeping in mind
that the question is for a college audience with existing
prior knowledge. Once all of the experts are done
reasoning, share an agreed conclusion.

Question text: {question_text}
Correct answer: {answer_text}”””

“““Based on the reasoning from these three experts and
their conclusion, reword these five points to begin with
action words from Bloom's Revised Taxonomy.
Reasonings: {reasonings}”””

“““Reasonings: {reasonings}
Five points: {points}

Of these five points, which one is the most relevant to
the question?”””
```

Figure 12.1: Three prompts used for the simulated expert prompting strategy for KC generation.

```
“““Below there is a multiple-choice question intended for
a {context} audience with existing prior knowledge on the
subject of {subject}. The question is used as a low-stakes
assessment as part of an {context} {subject} course that
covers similar content. The first answer choice, option
A), is the correct answer. If this question was presented
in a textbook for an {context} {subject} course, what five
domain-specific low-level detailed topics would the page
cover? Note that the question is for a college audience
with existing prior knowledge in {subject}.

Question text: {question_text}
{options_text}”””

"Based on these topics, reword them to begin with action
words from Bloom's Revised Taxonomy, while keeping them
domain-specific, low-level, and detailed."

"Of these topics, which is the most relevant to the
question?"
```

Figure 12.2: Three prompts used for the simulated textbook prompting strategy for KC generation.

In the first prompting strategy following the *simulated expert* approach, we employed the tree-of-thought technique, directing the LLM to emulate a discussion among three expert instructors [137]. The objective was to identify five specific, detailed skills and knowledge assessed by the provided MCQ. After this simulated discussion, the experts were expected to produce a list of the key skills they deemed necessary for answering the question. Subsequently, we introduced a prompt that instructed the LLM to refine the language of this list, aligning it with the verbs typically used in Bloom’s Revised Taxonomy [119]. This step was deliberately conducted after the initial list creation to avoid predisposing the selection of certain verbs, which we found reduced the quality of the labels during pilot testing. For instance, without this step, the experts would typically always suggest skills around the words of “understand”, “apply”, and “analyze” as they tried to strictly adhere to these levels of Bloom’s Revised Taxonomy. Finally, leveraging the insights from the discussion, the refined list of skills and knowledge aligned with Bloom’s Taxonomy, and the MCQ itself, the last prompt required the LLM to select the skills or knowledge most pertinent to the question at hand.

Our second prompting strategy, *simulated textbook*, drew inspiration from recent advancements in the creation of knowledge graphs, which frequently employ digital textbooks as a foundational source [40, 226]. These methods leverage structural elements like text headers or textbook indexes to outline the initial framework of the graph. Mirroring the expert approach, this strategy contextualizes the task around a MCQ as it might appear in a textbook. The LLM was tasked with identifying the specific, detailed topics a textbook page would cover if it included the given MCQ. Following this, a subsequent prompt directed the LLM to refine these topics, utilizing the language and verb categories found in Bloom's Revised Taxonomy [119]. The final step involved selecting the topic most relevant to the MCQ, ensuring that the identified knowledge areas were directly applicable to the question's context.

12.2.3 Human Evaluation

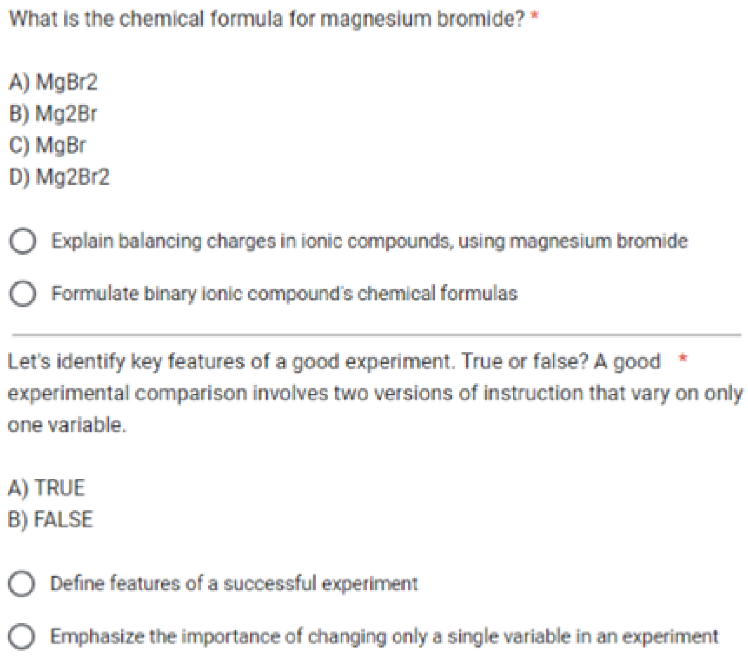
To benchmark this work, we initially compared the original KC tags, which were manually generated by the course creators, to those generated by the LLM. This comparison provided a baseline matching metric for each of the two prompting strategies across both domains. Beyond this direct match metric, our analysis extended to evaluating the outcomes of the second prompt within each prompting strategy. For both prompting strategies, this second prompt generates a list of the *top five* potential KCs for a MCQ, as identified by the LLM. This *top five* list was created with the intention of presenting it to an expert for selection, as part of potential future work. We considered it a partial success when the original manually generated KC tag appeared within this *top five* list, for either of the strategies. This occurrence was documented as a secondary, albeit less precise, metric of matching accuracy.

Recognizing that manually created KCMs can have issues, such as incorrect labeling or wording that misrepresents the required knowledge or skills, we implemented a secondary human evaluation. This aimed to assess the preference between LLM generated KCs and the original human-generated KCs, specifically for instances of mismatch. This secondary evaluation was only done for the mismatches from the simulated textbook strategy, as that had the highest matching percent to the original KC labels across both domains. For example, in a Chemistry MCQ, the human-generated KC was labeled "Use Gay Lussac's law" whereas the LLM-generated KC was “Understand gas pressure-temperature relationship". Given this discrepancy, we asked multiple human evaluators to choose which KC they believed more accurately matched the MCQ.

For this evaluation, we enlisted three domain experts in Chemistry and three in E-Learning. In the case of Chemistry, experts holding a bachelor's degree in Chemistry from the United States were recruited through Prolific, an online research platform [177]. These experts were given a survey implemented via Google Forms that tasked them with reviewing a series of MCQs accompanied by two KC labels. They were instructed to select the label that best matched a set of predefined KC criteria aligned with the KLI framework [116]. These criteria emphasized clarity, direct relevance to the subject matter, factual accuracy, and the ability to apply or integrate the knowledge into broader contexts or practical situations. Specifically, for the 35 MCQs identified as mismatches from the *simulated textbook* strategy in Chemistry, experts evaluated which of the two labels best met these criteria in relation to the MCQ. On average the task took roughly 28 minutes and participants received a compensation of $10.

The same procedure was applied to the E-Learning dataset, adhering to identical instructions and task formats as used for Chemistry. However, the three domain experts involved in this evaluation were instructional staff who had both participated in the E-Learning course and contributed to its development yet had not participated in the original creation of the KCM. For the ELearning dataset, there were 52 MCQs identified as mismatches based on the *simulated textbook* strategy. On average the task took roughly 32 minutes and participants received a compensation of $15.

An example of a Chemistry and E-Learning question used in both tasks can be seen in Figure 12.3. To ensure objectivity, the sequence of questions and the presentation order of the two labels for each question were randomized for each expert. To address the issue of LLM-generated KC labels being more verbose than their human-generated counterparts, we utilized the `gpt-4-0125-preview` API to refine these labels for clarity and brevity. We issued a prompt directing the LLM to rephrase each KC, ensuring the revised length did not exceed 1.5 times the word count of the corresponding human-generated KC. For example, a human generated KC consisting of ten words would result in an LLMrevised KC limited to a maximum of fifteen words. This approach was implemented to equalize the articulation level across labels and align with the typically concise format of KC labels. The preferred label was determined based on a majority vote, where at least two out of three experts had to agree on the choice. This methodological approach aimed to rigorously assess the comparative quality and applicability of LLM-generated KCs against those originally crafted by humans.



What is the chemical formula for magnesium bromide? *

A) MgBr2
B) Mg2Br
C) MgBr
D) Mg2Br2

○ Explain balancing charges in ionic compounds, using magnesium bromide

○ Formulate binary ionic compound's chemical formulas

Let's identify key features of a good experiment. True or false? A good experimental comparison involves two versions of instruction that vary on only one variable. *

A) TRUE
B) FALSE

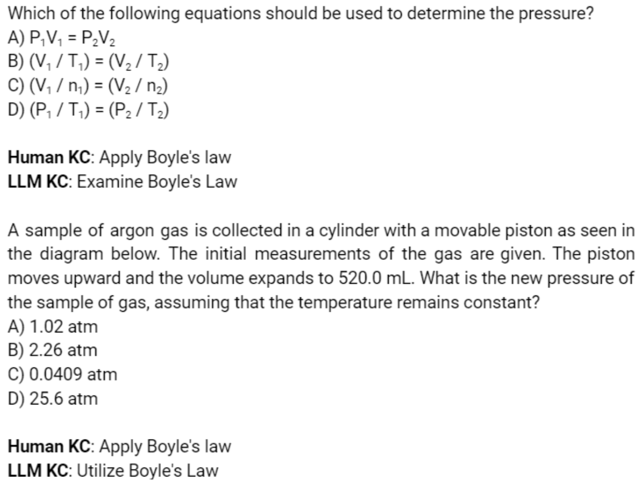
○ Define features of a successful experiment

○ Emphasize the importance of changing only a single variable in an experiment

Figure 12.3: A Chemistry MCQ (top) and E-Learning MCQ (bottom) used in this study.

12.2.4 Generating KC Ontologies

The prompting strategies described above can be employed to generate KCs for each individual question. One limitation of this methodology is that two questions that assess the same KC can receive LLM-generated labels that both contain the correct semantic information, but that feature two different wordings, as seen in Figure 12.4. Thus, the resulting KCM might contain redundancies which need to be resolved before using the KCM for purposes of assessing students' KC mastery or problem sequencing.



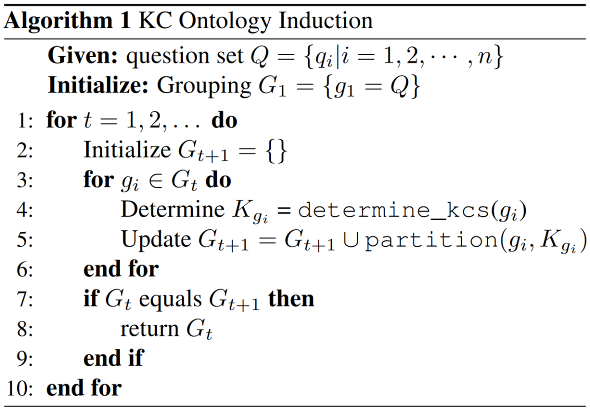
Which of the following equations should be used to determine the pressure?
A) $P_1V_1 = P_2V_2$
B) $(V_1 / T_1) = (V_2 / T_2)$
C) $(V_1 / n_1) = (V_2 / n_2)$
D) $(P_1 / T_1) = (P_2 / T_2)$

Human KC: Apply Boyle's law
LLM KC: Examine Boyle's Law

A sample of argon gas is collected in a cylinder with a movable piston as seen in the diagram below. The initial measurements of the gas are given. The piston moves upward and the volume expands to 520.0 mL. What is the new pressure of the sample of gas, assuming that the temperature remains constant?
A) 1.02 atm
B) 2.26 atm
C) 0.0409 atm
D) 25.6 atm

Human KC: Apply Boyle's law
LLM KC: Utilize Boyle's Law

Table 12.4: Two Chemistry MCQs targeting the same KC, with slight wording variations by the LLM.



Algorithm 1 KC Ontology Induction

```
Given: question set $Q = \{q_i | i = 1, 2, \cdots, n\}$
Initialize: Grouping $G_1 = \{g_1 = Q\}$
for $t = 1, 2, \ldots$ do
    Initialize $G_{t+1} = \{\}$
    for $g_i \in G_t$ do
        Determine $K_{g_i}$ = $\texttt{determine\_kcs}(g_i)$
        Update $G_{t+1} = G_{t+1} \cup \texttt{partition}(g_i, K_{g_i})$
    end for
    if $G_t$ equals $G_{t+1}$ then
        return $G_t$
    end if
end for
```

To promote alignment between the generated KCs, we propose an algorithm that induces an ontology of KCs of increasing granularity by iteratively partitioning the question pool into multiple groups. The overall algorithm is presented in Algorithm 1. The algorithm employs two prompts, shown in Figure 12.5, fulfilling two distinct tasks: (i) Determine a set of learning objectives that can be used to partition the question pool; (ii) Assign each question to one of the learning objectives to form groupings. After grouping the questions, the algorithm uses recursion and continues partitioning the individual subgroups until the LLM learning objectives are of finer granularity or until a group only contains a single question. We phrase the task of identifying KCs as determining fine grained learning objectives to be more aligned with common language. While it seems tempting to directly employ the partitioning induced by the first prompt, we noticed that when working with large questions list (e.g., >50) GPT-4 can fail to execute the instructions correctly either omitting questions in the assignment process or by assigning the same question to multiple groups. Having an explicit classification prompt that assigns each question to the most relevant group resolves this issue.

```
DETERMINE KCS PROMPT """Below there is a list of questions
and answers intended for a {context} audience with
existing prior knowledge on the subject of {subject}. You
are an educator who sorts the questions based on learning
objectives into groups. Ensure that each question belongs
to EXACTLY one group (not more or less).

Question List:
{question_list}
```

```
Use the following output format:
Group 1 name: [learning objective]
Group 1 questions: [Q1_1, ..., Q1_j]
...
Group N name: [learning objective]
Group N questions: [QN_1, Q_N_k]“““

CLASSIFY QUESTION PROMPT “““Below there is a question, its
answer, and a list of learning objectives. You are a
{subject} educator that determines the learning objective
that is most relevant to the question.

Question: {question}

Learning Objectives:
{objectives}

Use the following output format:
Most relevant Objective: [OBJECTIVE NUMBER]“““
```

Figure 12.5: Two prompts are used to determine appropriate learning objectives and then to classify the individual questions.

The iterative partitioning of questions induces an ontology of KCs of increasing levels of granularity. The expert can then decide which level of granularity is most suitable for their application. In this work, we employ the expert labeled datasets to evaluate the quality of our groups. Each dataset features a set of KCs denoted K and a set of questions denoted as Q. By design each $k \in K$ is associated with two questions $q_{k,1}$, $q_{k,2} \in Q$. A question grouping G is characterized by a set of disjoint groups {g_1, … , g_n} each hosting a subset of Q (i.e., $g_i \subseteq Q$). The optimal grouping G^* is characterized by |G^*| = |K| and for all $k \in K$ there is an $i \in$ 1, … , |K|, such that gi = {$q_{k,1}$, $q_{k,2}$}. Each step of our algorithm induces a grouping G_t . To assess the quality of these groupings at each step we define grouping accuracy and grouping refinement measures as follows:

$$\mathrm{acc}(G) = \frac{1}{|K|} \sum_{k \in K} 1\left[\exists g_i \in G : q_{k,1} \in g_i \wedge q_{k,2} \in g_i\right] \tag{1}$$

$$\mathrm{ref}(G) = \frac{1}{|Q|} \sum_{g_i \in G} \frac{|g_i|}{|\{k \,:\, \exists q \in g_i, \text{ s.t. } KC(q) = k\}|} \tag{2}$$

Grouping accuracy describes the proportion of question pairs which are correctly co-located in one of the groups. Grouping refinement evaluates the average number of KCs in each group. The optimal grouping G^* has an accuracy of 1 and a refinement of 1. The initial dataset that hosts all questions in a single set has an accuracy 1 and refinement 1 ⁄ $|K|$. We want our algorithm to increase the refinement of the groupings while maintaining a high level of accuracy (i.e., we do not want to split up question pairs of the same KC).

Table 12.1 displays the token counts and costs associated with two KC generation prompting techniques, expert and textbook, as well as the generation of KC ontologies for each domain.

Method	MCQ Count	Total Tokens	Prompt Tokens	Completion Tokens	Cost
expert	160	462,880	307,680	155,200	8.00
textbook	160	436,480	239,040	193,120	8.00
chemistry ontology	80	104,548	97,736	6,812	3.34
e-learning ontology	80	87,742	81,114	6,628	2.83

Table 12.1: Summary of token distribution and associated costs for the different prompting approaches.

12.3 Results

In our study, we initially evaluate the effectiveness of the two prompting strategies within the domains of Chemistry and ELearning. This evaluation is based on how well each strategy's outcomes align with the expert KCM. Subsequently, we explore the preferences of domain experts for the KC labels when discrepancies arise, determining whether they favor labels generated by human experts or those produced by the LLM. Lastly, we examine the performance of our ontology induction algorithm in both domains, focusing on its capability to categorize unlabeled questions by identifying shared KCs.

12.3.1 KCM Match Success

For the first part of this study, we evaluated how well the KCs generated by the LLM aligned with the KCs originally assigned to MCQs by their authors, across two distinct prompting strategies. Our assessment included a direct comparison of the LLM generated KC to the author-assigned KC for each MCQ. Additionally, we examined the top five KCs proposed by the LLM from the second prompt in both strategies to determine if any of these suggestions matched the author-assigned KC. Furthermore, we explored whether each MCQ was correctly categorized by only one strategy or if both strategies successfully identified the correct KC. The outcomes of these comparative analyses are presented in Table 12.2.

Method	Chemistry		E-Learning	
	Expert	Textbook	Expert	Textbook
Direct Match	42/80 (52%)	45/80 (56%)	28/80 (35%)	28/80 (35%)
Top Five	64/80 (80%)	63/80 (79%)	45/80 (56%)	50/80 (63%)
Matched Exclusively	9/80 (11%)	12/80 (15%)	9/80 (11%)	9/80 (11%)
Matched by Both	33/80 (41%)		19/80 (24%)	

Table 12.2: For each domain (Chemistry & E-Learning) and strategy (Expert & Textbook), the performance of LLM-generated KCs in relation to the existing KCM. The frequency of direct matches with the human tagged KC; instances where the KC was present in the top five LLMgenerated KCs; occasions where a KC was uniquely identified by only one strategy; and cases where both strategies matched the human tagged KC.

Our two-proportion z-test comparing the KC match rates of the *simulated textbook* strategy for Chemistry (42/80, 52%) and ELearning (28/80, 35%) questions revealed a significant difference (*Z=2.698, p=.007*). This indicates a statistically significant better performance of the *simulated textbook* strategy for Chemistry over E-Learning at p < .05, rejecting the null hypothesis of no difference in KC match rates.

We further explored the effectiveness of the *simulated textbook* strategy for identifying KCs across the MCQs in both domains. This analysis focused on the 40 KCs in each domain, where each KC was linked to 2 MCQs, to assess the accuracy of their tagging. In the Chemistry domain, the *simulated textbook* strategy successfully matched both MCQs to their correct KCs in 15 out of 80 cases (19%), correctly matched just one of the two MCQs also in 15 out of 80 cases (19%), and failed to match the KC in either MCQ for 10 out of 80 cases (13%). Similarly, in the E-Learning domain, both MCQs were accurately matched with their KC in 7 out of 80 cases (9%), only one of the two MCQs was correctly matched in 14 out of 80 cases (18%), and both MCQs failed to be matched in 19 out of 80 cases (24%). These results reveal the LLM's variable success rate in precisely identifying KCs through MCQs across different educational domains, suggesting superior performance for Chemistry compared to E-Learning. However, a chi-square test of independence revealed no significant association between the domain and the three aforementioned matching outcome categories (*X^2 (2, N=160) = 5.737, p=.057*).

12.3.2 Human KC Preference

For the MCQs in both the Chemistry and E-Learning domains that did not have a successful match with their KC using the *simulated textbook* strategy, we established the preferred KC label through the consensus of three domain expert human evaluators. A KC label was considered preferred only if at least two out of the three evaluators agreed on its selection. From the *simulated textbook* strategy, Chemistry had 35 MCQs where the human and LLM KCs were mismatched, and E-Learning had 52 MCQs.

Within the Chemistry domain, analysis of 35 MCQs revealed a clear preference for the LLM-generated KC labels, which were chosen in 23 out of 35 cases (66%), compared to human-generated labels preferred in 12 instances (34%). Additionally, we observed a substantial level of agreement among the experts, with two-thirds majority agreement (at least two evaluators in agreement) occurring in 25 out of 35 cases (71%), while unanimous agreement (all three evaluators in agreement) was found in 10 cases (29%). Similarly, in the E-Learning domain, upon examining 52 MCQs, LLM-generated labels were preferred in 32 cases (62%), with human-generated labels being chosen in 20 cases (38%). The evaluators demonstrated a clear consensus, with two-thirds majority agreement present in 34 out of 52 instances (65%) and unanimous agreement observed in 18 instances (35%). A comparison of the preferences by domain can be seen in Figure 12.6. Aggregating preference data from both domains we can verify a statistically significant preference for the LLM-generated KCs. A two-sided binomial test was conducted to assess whether the human evaluators exhibit a preference towards expert or LLM generated KC labels. For 57 out of 87 evaluated MCQs, the evaluators favored the LLM-generated labels, indicating a statistically significant preference (*p=0.017*).

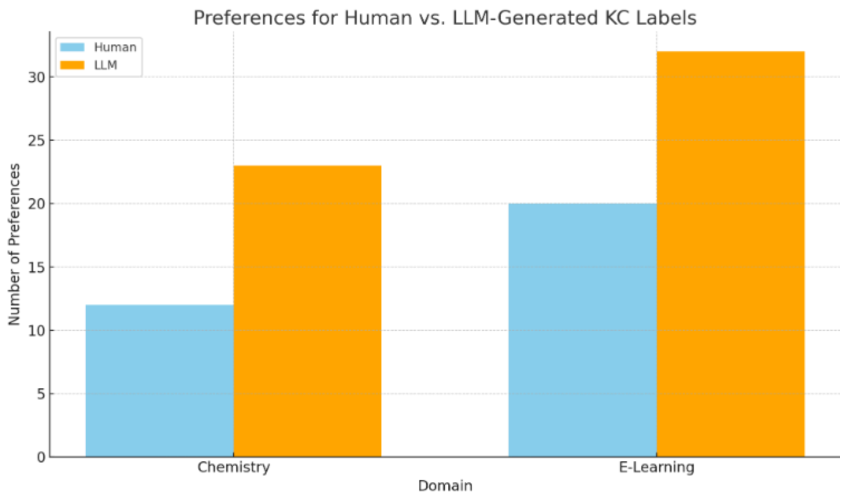


Figure 12.6: Comparison of domain expert preferences for human- vs. LLM-generated KC labels.

12.3.3 Generated KC Ontologies

We now focus on the KC ontologies generated by the clustering algorithm for the Chemistry and the E-Learning datasets. An excerpt of the KC ontology for Chemistry is shown in Figure 12.7. Going from the root of the tree downwards we can observe how the KCs identified by the algorithm increase in granularity at each step.

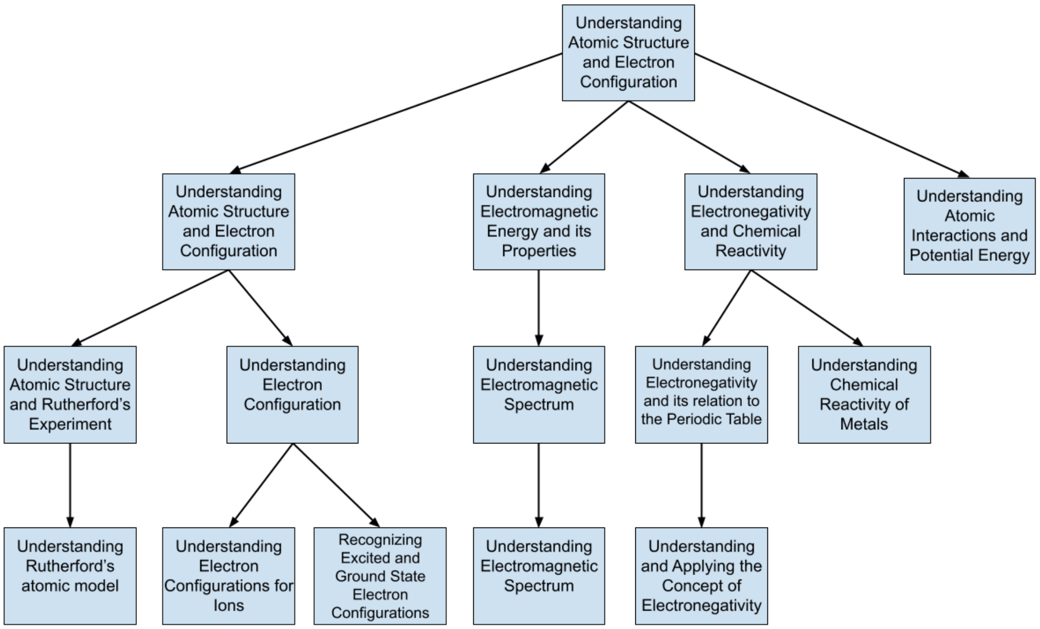

Understanding Atomic Structure and Electron Configuration
Understanding Atomic Structure and Electron Configuration
Understanding Electromagnetic Energy and its Properties
Understanding Electronegativity and Chemical Reactivity
Understanding Atomic Interactions and Potential Energy
Understanding Atomic Structure and Rutherford's Experiment
Understanding Electron Configuration
Understanding Electromagnetic Spectrum
Understanding Electronegativity and its relation to the Periodic Table
Understanding Chemical Reactivity of Metals
Understanding Rutherford's atomic model
Understanding Electron Configurations for Ions
Recognizing Excited and Ground State Electron Configurations
Understanding Electromagnetic Spectrum
Understanding and Applying the Concept of Electronegativity

Figure 12.7: A section of the tree structure demonstrating KC ontology refinement for part of Chemistry.

To evaluate the quality of the KCM at different steps we employ the grouping *accuracy* and *refinement* metrics defined in Section 3.4. First, for Chemistry (Figure 12.8, left) the algorithm converges within 6 iterations to a KCM that groups the 80 questions into 42 different KCs–close to the expert model with 40 KCs. At time of convergence the grouping accuracy indicates that 65% of the question pairs are matched correctly and the grouping refinement of 0.804 indicates that the majority of nodes only feature questions belonging to a single KC. Second, for E-Learning (Figure 12.8, right) the algorithm converges within 5 iterations to a KCM that groups the 80 questions into 63 different KCs-exceeding the expert model with 40 KCs. At the time of convergence, the grouping accuracy is 17.5% and the grouping refinement is 0.848. Because the final ELearning KCM employs 63 KCs, many KCs are only tagged to a single question leading to splits between the expert defined question pairs explaining the low accuracy. This suggests that LLM-generated KCM is of finer granularity than the human expert KCM. In real world applications, the domain expert might want to employ a lower level of KC granularity which can be achieved by terminating the algorithm early.

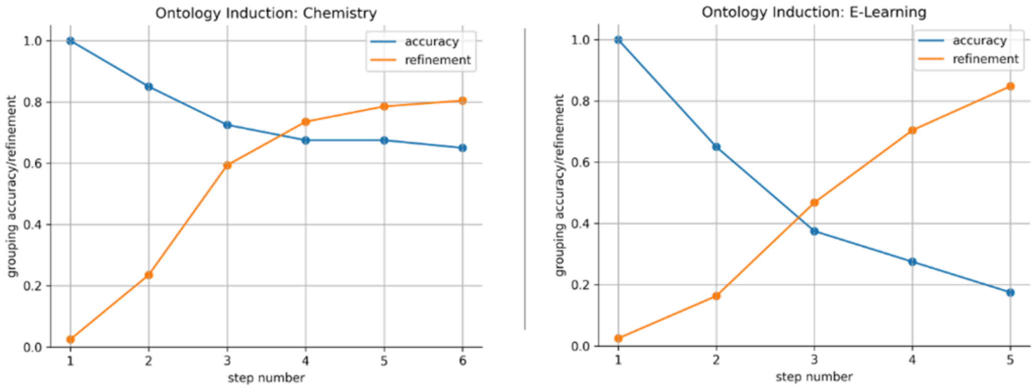


Figure 12.8: Information about the 30 questions used in this research.

12.3.4 KC Ontology Verification

To evaluate the effectiveness of the ontology grouping method, we conducted a comparative analysis against other straightforward KC grouping strategies. Specifically, we compared the results of three different steps in our ontology method with the following groupings:

Expert Model: The 40 KCs originally assigned to the 80 MCQs by their author.

All Model: Treating all questions as a single KC, assuming that they all assess the same underlying skill or knowledge.

Single Model: Assigning each individual question to its own unique KC, maximizing the granularity of the grouping.

To ensure a fair comparison, we used the item-blocked cross-validated Root Mean Square Error (RMSE) as our evaluation metric [90]. This approach involves partitioning the data based on items (MCQs) rather than students, which helps in assessing how well the model predicts unseen questions. We focused on the first three iterations of the ontology-based grouping method to analyze how incremental refinements impact the model's predictive accuracy. Each iteration represents a stage in the refinement of the ontology, where each step is further refinement, meaning the MCQs are put into more groups corresponding to KCs. By examining these steps, we aimed to determine whether successive improvements in the ontology lead to better model performance compared to simpler grouping methods, while also balancing a realistic use case of trying the first three iterations, thus reducing time and cost.

Table 12.3 compares the performance of the first three ontology grouping steps with the All, Single, and Expert models for the 80 MCQs in both Chemistry and E-Learning. Using student data collected from the fall 2022 courses, we evaluated each model's performance using item-blocked RMSE, which measures the average difference between predicted and actual values, with lower values indicating better model accuracy.

For Chemistry, we analyzed 3,071 observations, while for E-Learning, there were 4,130 observations. In the Chemistry domain, the Step 2 ontology model had the best performance, whereas in E-Learning, the Expert model performed the best. Across both domains, all three ontology-based models (Steps 1–3) outperformed the All and Single models. Notably, in E-Learning, the Step 3 and Expert models had nearly identical performance, differing by just 0.0005 RMSE, demonstrating highly comparable results.

Model	Chemistry KC #	Chemistry RMSE	E-Learning KC #	E-Learning RMSE
Step 1	30	0.4281	22	0.3729
Step 2	43	**0.4224**	31	0.3716
Step 3	45	0.4244	38	0.3668
Expert	40	0.4273	40	**0.3663**
All	130	0.4989	135	0.5103
Single	1	0.4464	1	0.3874

Table 12.3: Comparison of different KC models for Chemistry and E-Learning, showing the number of KCs assigned to the 80 MCQs and their item-blocked cross-validated RMSE.

12.4 Discussion

Our results demonstrate the potential of leveraging LLMs to generate and assign high-quality KCs to educational questions. Specifically, in the undergraduate Chemistry domain, we successfully matched more than half of the MCQs with their corresponding KCs, and in the master's E-Learning domain, we achieved a match rate of one-third. For MCQs whose KCs were not directly matched by the LLM, domain expert evaluations showed a two-thirds preference for LLM-generated KCs over the existing human-generated alternatives. Additionally, we introduced a novel clustering algorithm for grouping questions by their KCs in the absence of explicit labels. These results propose a scalable solution for generating and tagging KCs for questions in complex domains without the need for pre-existing labels, student data, or contextual information.

The higher match rate for Chemistry compared to E-Learning can potentially be traced back to the quality of the human-generated KCM. The Chemistry KCM featured more specific KCs, each typically incorporating just a single piece of domain-specific jargon, unlike the broader KCs with multiple terms found in the ELearning KCM. Additionally, introductory Chemistry topics are likely to be more prevalent in the LLM's training data than the specialized E-Learning content, which might have contributed to this discrepancy. Despite reasonable success in identifying the top five KCs, the LLM faced difficulties in accurately selecting the most appropriate KC. It often favors general options over precise and domain-specific ones, potentially due to the presence of domain jargon. This led to a substantial portion of MCQs in both domains, 21% in Chemistry and 38% in E-Learning, not matching any of the top five KCs suggested by the LLM. These results indicate that while LLMs are capable of surfacing relevant KCs, the specific nature of domain jargon and a bias towards generalization can impede the accurate identification of a KC.

We observed a notable and statistically significant (two-thirds) preference for LLM-generated KC labels over human-generated ones in both domains. This is possibly due to their slightly longer length and the enhanced readability afforded by the LLM's advanced next-word prediction capabilities. Given this preference, it might suggest the importance of prioritizing human evaluative feedback over direct matches with existing KCMs when assessing LLM effectiveness in generating KCs, especially considering the subjective nature of KC evaluations which can vary significantly based on the reviewer's perspective [134]. Another potential explanation may be that the original KCM designed by the course authors was imperfect, containing KCs that did not fit the particular MCQ or that were too broad. Interestingly, despite the inherent subjectivity in KC evaluation, all six reviewers in our study showed a preference for LLM labels. These findings highlight a pronounced preference for LLM-generated KC labels over human-generated ones across both domains. The higher preference for LLM labels, along with the levels of agreement among evaluators, suggests that LLM-generated labels could serve as an effective substitute for manually created labels in categorizing MCQs by their KCs. This preference does not imply that LLM-generated labels should completely replace human input; rather, they could at least provide a valuable foundation, enabling human experts to further select or refine the KCs identified by the LLM. This collaborative approach leverages the strengths of both LLM capabilities and human expertise, potentially leading to more accurate and universally acceptable KC categorizations.

When generating KCs for pairs of questions that assess the same KC, we found that the LLM can assign labels with the correct semantic information, but with different wordings (e.g., see Figure 4). To resolve these redundancies in the KCM, we proposed and evaluated an algorithm that iteratively partitions the question pool to generate KCMs of increasing granularity. The KC ontology induced by this process is similar to taxonomies such as the Common Core State Standards [194]which allow for the categorization of learning materials at different levels of specificity (refer to Figure 7). For the Chemistry questions, we observed that the KCM at the convergence of the algorithm is of similar granularity as the expert model and most expert identified question pairs are grouped correctly. For the E-Learning questions, the converged LLM-generated KCM featured significantly more KCs than the expert model (63 vs 40) indicating a higher granularity. Because of this, our evaluation metrics–which were grounded in the expert KCM–assigned the LLM-generated KCM a low grouping accuracy. Based on the human evaluation of human and LLM-generated KCs, this might indicate that the expert KCM for the E-Learning course contains inaccuracies and is of lower quality. Lastly, LLM induced KC ontologies might support domain experts structure subject content and provide them with control over the level of KC granularity that is most appropriate. After deciding on a set of KCs the resulting taxonomy could provide a foundation for other types of automated KC tagging algorithms (e.g. [78, 180, 182]).

Given the preference for LLM-generated KC labels observed in the human evaluations across both domains, practitioners could consider using these labels as initially provided. However, a more effective approach would involve implementing a human-in-the loop system, where domain experts review and confirm the appropriateness of these labels, creating their own alternatives if necessary. Ideally, this process would start with the LLM generated labels being preliminarily assigned to problems, followed by a verification step where experts could either approve, modify, or replace them as needed. This process not only ensures accuracy, but also significantly reduces the time and effort required compared to starting from scratch. Ultimately, while the initial LLM-generated labels serve as an effective preliminary pass in developing a knowledge component model, they should be seen as a foundation that can be further refined based on expert insights and student performance data [94].

12.5 Limitations & Future Work

In our research, we introduced innovative methods for generating and grouping KCs using a LLM. However, this approach is subject to certain limitations, such as the opaque nature of LLMs, their susceptibility to unexpected output variations, and the potential for biased results [190]. To address these challenges and improve the reliability and efficiency of our methods, we employed a specific iteration of GPT-4, accessed through the `gpt-4-0125-preview` API. This strategy was designed to standardize the evaluation process and guarantee the reproducibility of results by producing consistent outputs in response to predefined prompts. Despite these efforts, the choice of wording in prompts remains a critical factor, significantly affecting the model's output due to LLMs' inherent sensitivity to input nuances. Moreover, the process of evaluating KC quality is complicated by human subjectivity, even among domain experts following specific and detailed guidelines. The definition of a "good" KC is still not clear-cut [116], as reflected in prior literature discussing the desired granularity, making the evaluation process heavily dependent on individual judgment. Our study's scope was limited to two domains, restricted by the scarcity of suitable datasets. Our attempts to use datasets similar to those in prior studies were obstructed due to their unavailability for access. Additionally, niche domains may exhibit poorer performance and higher inaccuracies due to their limited representation in the LLM's training data. For example, this limited representation could explain the low agreement between the expert-created KCM and the E-Learning KC ontology generated by the algorithm.

In future research, we aim to broaden the application of our methods to questions from additional domains and various formats, such as short-answer questions. We are interested in investigating the impact of different contexts, such as instructional text provided before a question, on the quality of KCs generated by LLMs. Our goal is to further refine the prompting strategies we have developed and to foster collaboration among researchers and educators by making our data and code publicly available. Additionally, we plan to explore the potential benefits of utilizing different LLMs, which may enhance the results or provide greater consistency, a notable challenge in current LLM education research.

12.6 Conclusion

KCs are crucial for modeling student learning and empowering educational technology with adaptivity and analytics. Therefore, simplifying and scaling the creation and association of KCs with educational content across various domains is essential. In this context, our study suggests that LLMs can play a significant role in facilitating this process. We developed a method to generate KCs for assessment items relying solely on the questions' context and demonstrated its success with assessments from Chemistry and ELearning courses. Our findings indicate that, although the direct matches between LLM-generated and human-generated KCs were moderate, domain experts most frequently preferred the LLM generated KCs for the assessments. To overcome the challenge of categorizing assessments by their underlying KCs without labels or context, we also introduced an algorithm for inducing KC ontology and clustering assessments accordingly. Despite the subjectivity, time, and domain expertise that is typically part of the KC mapping process, our approach represents a step towards a scalable solution that addresses these challenges across complex domains. Our research highlights the potential of LLMs to enable individuals, regardless of their technical skills or domain knowledge, to contribute to the development of Knowledge Component Models.

Chapter 13
Discussion and Future work

This thesis presents methods of crowdsourcing, learnersourcing, and utilizing NLP techniques to evaluate educational content and generate associated skills. Through my research, I contribute practical approaches for incorporating learnersourcing into classrooms, making it accessible to educators. I explored multiple methods for skill tagging educational content in an automated and scalable way. Additionally, I critically assessed human evaluation techniques for educational content and demonstrated how rubrics can enhance the evaluation process. Finally, I highlighted the limitations of existing question evaluation methods and proposed a more effective alternative.

In the following sections, I reflect on the implications and lessons learned from this research, and suggest avenues for further inquiry. My work aims to contribute to ongoing discussions in these areas and provide practical insights for future research.

13.1 Learnersourcing

Learnersourcing involves engaging students in the creation and evaluation of educational resources, which can also enhance their learning. While dedicated systems for learnersourcing exist, they are not always necessary to achieve high-quality responses and strong participation rates. Even without these systems, participation tends to be equitable and reflective of class demographics. While the generation of new resources and course improvements is valuable, it's essential to ensure that students retain autonomy and continue learning throughout the process.

13.1.1 Effectiveness of Learnersourcing

Learnersourcing MCQs and SAQs in low-tech and low-stakes environments is effective and easy to implement in most classes (*chapters 3 & 4*). Learnersourcing systems are highly effective, and students routinely use them to generate and evaluate a significant amount of educational content [168, 212]. However, even low-tech solutions like Google Forms with short answer text boxes or ones preferred learning management system can be just as effective [162]. Students can create high quality output from learnersourcing activities that involve the generation of MCQs, SAQs, hints, feedback, and skills using these simple tools. Even without the affordances offered by learnersourcing systems, such immediate feedback or advanced displays, students are fully capable of contributing high-quality responses to these learnersourcing activities.

It is important to consider lowering the barriers to participation and involving more stakeholders, including instructors, in the learnersourcing process. The entry barrier is quite low, and the benefits to student learning are substantial, while also providing valuable resources to improve courseware. If an instructor is already using a tool or ed-tech platform with short answer text boxes, they should consider trying a learnersourcing activity. They might find improvements for their course and discover something new, their students certainly will!

13.1.2 Equitable Participation

Confirms that equitable participation occurs in optional learnersourcing environments, with high-quality contributions and active involvement (*chapter 7*). Learnersourcing systems are often used as required assignments within courses, based on the assumption that optional activities would lead to low participation and poor-quality contributions [156, 162, 213]. However, our work has proven that this is not the case, as participation rates remain strong, and the contributions include a diverse mix of high-quality responses [158, 159]. A related, often overlooked aspect of research on learnersourcing or any optional educational activity is identifying who may be choosing not to participate. Our findings show that participation in our optional learnersourcing activities is representative of the entire course demographic.

Future research on optional activities, whether they involve learnersourcing or not, should consider reporting demographic information to ensure that no group of students is being overlooked. This is important because every student's unique perspective can enhance the value of the activity and the resources students contribute. This inclusion of diverse, novice perspectives is a key part of what makes learnersourcing so valuable and innovative.

13.2 Skill Tagging

Skill tagging educational content often relies on a mix of student performance data and human articulation, which can be a challenging and subjective process to replicate at scale. Recent research has focused on using automated ML and NLP methods for skill tagging, utilizing a wide variety of data sources. We explored a scalable approach using crowdsourcing and learnersourcing, but this proved to be less successful than existing methods and presented several challenges. In response, we leveraged LLMs for a scalable and automated method, and found this approach to be quite successful in the two domains we tested. However, further testing is necessary to validate these findings across more domains and against empirical student data.

13.2.1 Necessity of Domain Expertise

Provides further evidence that skill tagging may require domain knowledge or expertise, even with various scaffolding approaches (*chapters 5 & 6*). Previous research states that skill tagging can be done through human methods, such as Cognitive Task Analysis (CTA) performed on domain experts [26, 163]. Even think-aloud protocols with novices can provide insights into the skills involved in the problem solving process [203]. However, since these represent two different levels of expertise, we explored what range of novices, through crowdsourcing and learnersourcing, could accurately contribute to skill tagging if the task was heavily scaffolded. While the results for both crowdsourced and learnersourced approaches were less promising, there were some areas where they accurately identified certain skills.

This led us to reconsider how we might utilize the crowd and students in this process and how we could leverage the "knowledge" of LLMs to supplement or complement their contributions. We believe this is a promising direction for learnersourcing, particularly given its current success in areas like question generation and hint generation [60, 214].

13.2.2 New Method for Skill Tagging

We developed a new method for knowledge component generation and skill tagging for unlabeled multiple-choice questions (*chapter 12*). Building on our prior work and advancements in ML and NLP-based methods, we hypothesized that LLMs could offer an effective solution for skill tagging. Our findings confirmed that LLMs provide a quick, scalable, and accurate way to generate skill tags for higher education MCQs [166]. The process is straightforward, requiring only a set of questions and a series of prompts, without the need for additional context or metadata. Human evaluators favored the skill tags generated by LLMs, finding them more readable, fine-grained, and easier to understand.

We believe this focus on human-readable labels is crucial for the skill tags, yet often downplayed in previous research [21, 62]. While technical metrics like AIC/BIC scores are important, in certain contexts, they are less meaningful if the labels are not easily interpretable by instructors. For example, in a learning dashboard, the usability of the labels may be more critical than marginal improvements in model performance. This principle drives much of our current work, as we aim to develop tools that are not only effective, but also practical for instructors and learning engineers in classroom settings.

While we demonstrated success, further testing of this method is needed. Future research should explore its effectiveness with other LLMs, across different domains, and in empirical comparisons with existing models and student data. We plan to integrate this approach into SAQUET, enabling not only question evaluation, but also the generation of a set of hypothesized skills that each question assesses.

13.3 Human Expertise and Evaluation

In many research areas, particularly in NLP, human evaluation is considered the gold standard. This is especially true for automated question generation, where human evaluation has traditionally been the benchmark, often relying on the use of subjective scales. To address the need for a more standardized and consistent approach, we utilized the IWF rubric for MCQs and a 9-item rubric for SAQs. Our goal was to develop a scalable method for human evaluation via a crowdsourcing approach. We hypothesized that if the evaluation task was properly scaffolded and guided by a clear rubric, crowdworkers could effectively assess question quality. By standardizing the evaluation process, we aimed to create a more consistent, reliable, and scalable method for evaluating the quality of educational questions. This approach not only reduces subjectivity via verified rubric use, but also makes it possible to scale.

13.3.1 Crowdsourced Content Assessment

Validates that crowds can effectively assess the quality of content using a rubric (*chapters 8 & 11*). Providing the IWF rubric for MCQs and the 9-item rubric for SAQs to the crowd allowed them to evaluate questions across various higher education domains with a level of accuracy comparable to expert evaluation [155, 165]. Remarkably, even without domain-specific knowledge and potentially limited understanding of the content, the crowd could accurately apply these rubrics. However, setting up these crowdsourcing tasks proved to be time-consuming and not cost-efficient, which limited the scalability of this approach. Despite these challenges, the initial success demonstrated by the crowd suggests that this is a promising area for further exploration. Future research could focus on engaging more advanced crowdworkers and potentially enhancing their workflows with LLMs or other innovative methods to improve efficiency and scalability.

13.3.2 Advanced Domain Knowledge in Evaluations

Suggests that while students or non-experts can sometimes conduct evaluations, advanced domain knowledge may be necessary, potentially provided by LLMs (*chapters 8 & 11*). While we achieved promising results with minimal crowdsourcing filters when applying the IWF rubric to MCQs, our success was more limited with SAQs. When comparing the performance of LLMs to that of the crowd, we found that a more advanced crowd, specifically individuals with bachelor's degrees in the relevant domain recruited via Prolific, achieved the highest accuracy in applying the rubrics.

This suggests that human evaluation, particularly by individuals with domain-specific knowledge, remains as one of the most comprehensive methods for question evaluation. Leveraging a knowledgeable crowd can lead to superior results across both MCQs and SAQ, highlighting the importance of expertise in the evaluation process.

13.3.3 Human-AI Hybrid Evaluation

Introduces a human-AI hybrid approach for the evaluation of questions (*chapter 11*). Building on our prior crowdsourcing work for content generation and evaluation, we recommend a hybrid approach that leverages both LLMs and human expertise. In this approach, LLMs would handle the simpler rubric criteria where they have demonstrated high accuracy, while humans, through crowdsourcing, learnersourcing, or other methods, would focus on the more complex criteria that still require human judgment. Drawing from prior research, this approach could involve LLMs taking an initial pass at the evaluation, with humans subsequently reviewing and either accepting or rejecting the LLM's recommendations [108]. This method has the potential to streamline the evaluation process, enhancing scalability without sacrificing quality.

The growing trend of human-AI hybrid applications, where LLMs and humans collaborate to refine outputs, supports the viability of this approach [60, 214]. In the context of question evaluation and generation, this method not only produces high-quality results but also promotes student learning, as demonstrated by previous studies [215].

13.4 Question Quality and Evaluation

Crowdsourcing, learnersourcing, and human-AI hybrid approaches all show potential for evaluating the quality of educational questions. However, we aimed to develop a fully automated, scalable method that goes beyond surface-level analysis and is grounded in learning sciences research. Our goal was to create a domain-independent approach that leverages not only LLMs, but also existing NLP techniques. Much of the existing literature on question evaluation relies on subjective assessments that are difficult to replicate, and the commonly used automated metrics were not originally designed to assess question quality, particularly in educational contexts. We set out to test these existing metrics and propose our own, offering a robust solution that can be applied to both new and existing questions by any user, regardless of their expertise.

13.4.1 Ineffectiveness of Existing Methods

Provides proof that commonly used methods for evaluating educational multiple-choice questions are inadequate (*chapter 10*). We applied five commonly used metrics for evaluating question quality to our dataset of MCQs tagged with the IWF criteria [154]. We found that questions with clear flaws sometimes scored better than flawless questions according to these metrics.

This outcome was not entirely unexpected, as these metrics primarily focus on aspects like word choice, readability, and reading level. For example, a question might score high in lexical diversity due to its use of unique and complex words, yet still contain implausible distractors, making it less effective as an assessment tool.

We do not fault researchers and practitioners for relying on these metrics, as they are among the few automated options available [170]. However, our findings make it clear that these metrics do not effectively distinguish between low and high-quality questions in educational contexts. This realization led us to pursue a method grounded in learning sciences, one that relies on proven tools like the IWF rubric, while also introducing the standardization necessary for a reliable evaluative metric. Although our approach is not the only possible solution, as there are other rubrics and methods that can be used to evaluate question quality, it is crucial that any metric be specifically tailored for educational assessments. Such a metric must consider how questions will be used with students who are actively learning and possess their own test-taking strategies and abilities.

13.4.2 New Method for MCQ Evaluation

Proposes a new method for the evaluation of educational multiple-choice questions (*chapters 9 & 10*). While still in development, we deployed and utilized the Scalable Automatic Question Usability Evaluation Toolkit (SAQUET), which demonstrated expert-level accuracy in applying the 19-criteria IWF rubric to higher education questions across various domains. Initially, SAQUET was purely rule-based (see *chapter 9*), utilizing a variety of NLP methods without incorporating LLMs. However, it has since evolved into a hybrid approach that integrates LLMs to handle cases where rule-based methods struggle to make confident decisions.

SAQUET is user-friendly and specifically designed to assess the quality of MCQs in terms of their pedagogical effectiveness. Our goal is to continue developing SAQUET into a comprehensive tool that allows users to quickly upload or create questions, receive a full evaluation, and even generate skill tags based on the work presented in *chapter 11*. We also envision users contributing back to SAQUET, as it is open-source, enhancing the methods and improving the accuracy and richness of its evaluations. There are still areas in question evaluation that require further research, such as ensuring alignment with learning objectives and associated content. As SAQUET evolves, addressing these aspects will be crucial to its success and utility in educational settings.

13.5 Future Directions

For future work, we intended to provide further empirical and data-driven evidence that these methods, both the skill tagging approach and SAQUET, are effective. We plan to continue testing across different domains using real student data to validate and refine our techniques, comparing our evaluation against data-driven approaches such as those from IRT. In areas where LLMs are utilized, we will focus on improving prompts with the assistance of collaborators and experimenting with different models as they evolve. We invite educators, researchers, and practitioners to engage with our work by offering their insights and improvements to further refine the criteria, as we have done. This form of collaboration would contribute to developing a more educationally robust metric enriched by collective expertise, while also providing invaluable feedback on the usability of the methods. Finally, building on our recent work, we aim to construct a tool that offers feedback from SAQUET and provides skill tag suggestions for MCQs. Eventually, we plan to expand these methods beyond MCQs to include other forms of assessment, such as SAQs, and even instructional content.

13.5.1 Actionable Improvements

Currently, SAQUET identifies potential flaws in the provided MCQs, but does not offer feedback on how to rectify them. As a next step, we plan to leverage LLMs, even for criteria that are purely rule-based, to provide feedback and make corrections directly to the MCQs. In early testing, we applied SAQUET to questions from university-level Chemistry and Statistics courses, generating reports that identified specific flaws in each question. Instructors for the respective courses then made quick corrections based on these reports. Examples of improved Chemistry and Statistics questions can be seen in Figures 13.1 and 13.2, respectively.

What are the subatomic particles known as protons?

A) Positively charged subatomic particles.

B) Sum of electrons and neutrons.

C) Negative subatomic particles

D) Discovered by Ernest Rutherford.

What is the charge of subatomic particles known as protons?

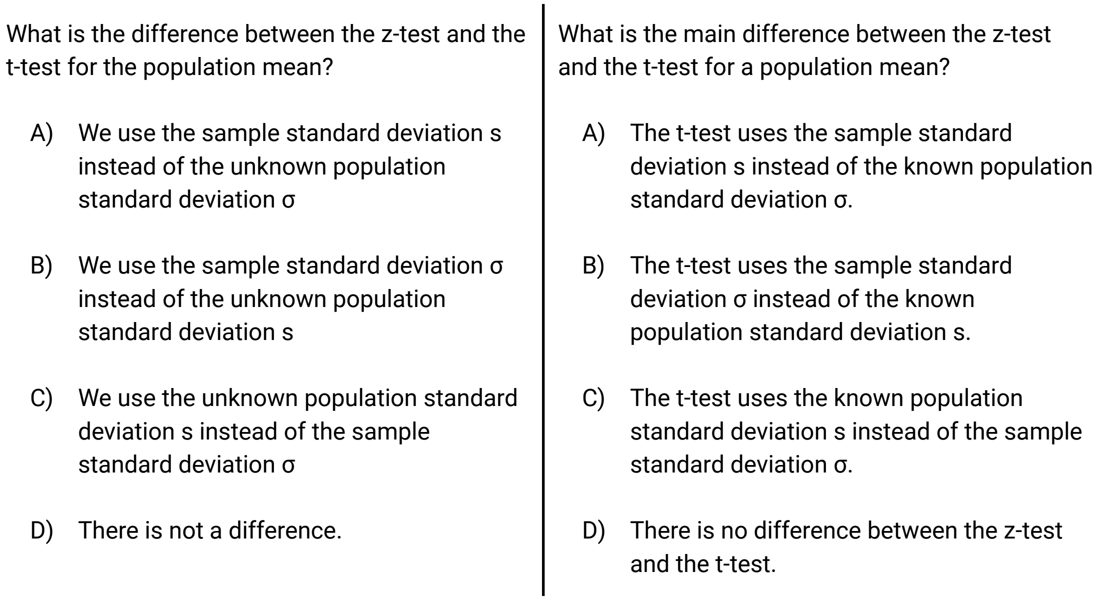
A) Positive.

B) Neutral.

C) Negative.

D) The combined charge of electrons and neutrons

Figure 13.1: On the left, a chemistry question with four IWFs: *grammatical cue*, *convergence cue*, *logical cue*, and *implausible distractor*. The improved version, revised by an instructor, is shown on the right.



What is the difference between the z-test and the t-test for the population mean?

A) We use the sample standard deviation s instead of the unknown population standard deviation σ

B) We use the sample standard deviation σ instead of the unknown population standard deviation s

C) We use the unknown population standard deviation s instead of the sample standard deviation σ

D) There is not a difference.

What is the main difference between the z-test and the t-test for a population mean?

A) The t-test uses the sample standard deviation s instead of the known population standard deviation σ.

B) The t-test uses the sample standard deviation σ instead of the known population standard deviation s.

C) The t-test uses the known population standard deviation s instead of the sample standard deviation σ.

D) There is no difference between the z-test and the t-test.

Figure 13.2: On the left, a statistics question with three IWFs: *convergence cue*, *grammatical cue*, and *implausible distractor*. The improved version, revised by an instructor, is shown on the right.

This feedback has proven effective in informing instructors, learning engineers, and instructional designers about potential flaws in the questions they create. However, to further enhance efficiency and question quality, we are working on having an LLM make the corrections automatically. While the system will still report the identified flaws and clearly outline the changes made to address them, allowing users to approve or reject the changes may streamline the process compared to making the revisions themselves.

13.5.2 Item-Response Theory

Much of the work in verifying the efficacy of SAQUET has relied on qualitative evaluations by domain experts, which, even with the use of a rubric, can introduce some subjectivity. To move toward more objective, data-driven approaches, we have recently begun exploring how different item flaws may relate to item difficulty and the prediction of item discrimination, using IRT. Item difficulty refers to the probability that a given test item will be answered correctly, with higher values indicating more challenging questions. Item discrimination measures how well an item differentiates between students of varying ability levels, with higher discrimination indicating that the item is more effective at distinguishing between high and low performers.

In our initial exploration, we used data from thousands of students working through STEM courses from a popular online learning platform. We found that some flaws, such as *logical cues* and *complex or k-type options*, were positively correlated with item difficulty, making questions harder. On the other hand, flaws like *all of the above* and *longest answer correct* were negatively correlated with difficulty, indicating they tend to make the questions easier.

Chapter 14
Conclusion

The rapid generation of educational content through crowdsourcing, learnersourcing, and generative AI offers significant potential. However, it often overlooks critical metadata, such as the specific skills being assessed by each question. Additionally, the evaluation techniques currently employed to ensure the quality of such content are inadequate and lack standardization. In my dissertation, I address these challenges by combining insights from Human-Computer Interaction and Learning Sciences, taking significant steps toward improving the quality of the educational content we give students.

My research began with an exploration of educational content creation through learnersourcing in low-stakes and low-technology environments. The findings demonstrate that high-quality educational content can be generated without the need for complex systems, and equitable participation is possible when students are empowered to make choices in the process. In the area of skill tagging, I explored human-in-the-loop methods that varied in domain knowledge and expertise, utilizing both crowdsourcing and learnersourcing. To address the shortcomings identified, I proposed new approaches that leverage LLMs to enhance the accuracy and efficiency of skill tagging content.

In both the creation and evaluation processes of educational content, I further examined the role of subject matter proficiency. My findings suggest that a hybrid approach, combining human insight with automated tools, could enhance these processes and scaffold the knowledge required. Current widely used evaluation methods often fall short of ensuring educational content quality and reliability, leading me to propose a new automated method that addresses these shortcomings, known as SAQUET. This new approach remains automated and scalable, while also focusing on the pedagogical aspects of the content.

This dissertation advocates for the development of mixed-initiative approaches that combine user input, whether from students, crowds, or domain experts, with generative AI capabilities to evaluate and create educational content. These interventions, along with future ones, should be designed to integrate seamlessly into existing digital courseware, ensuring accessibility without requiring complex systems or specialized knowledge. By leveraging existing learning science work and enhancing standardization,we can create educational technologies and content that are not only effective but also widely usable by instructors and learning engineers. Looking ahead, it is crucial to maintain the integration of proven learning science methods at the core of educational technology and interventions, especially in the area of human-AI collaboration. Ultimately, the goal is to continually improve learning outcomes for students.

Bibliography

[1] Abdi, S., Khosravi, H. and Sadiq, S. 2020. Modeling learners in crowdsourcing educational systems. *International Conference on Artificial Intelligence in Education* (2020), 3–9.

[2] Abdi, S., Khosravi, H., Sadiq, S. and Demartini, G. 2021. Evaluating the quality of learning resources: A learnersourcing approach. *IEEE Transactions on Learning Technologies*. 14, 1 (2021), 81–92.

[3] Addlesee, A., Sieińska, W., Gunson, N., Garcia, D.H., Dondrup, C. and Lemon, O. 2023. Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering. *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (2023), 229–241.

[4] Aflalo, E. 2021. Students generating questions as a way of learning. *Active Learning in Higher Education*. 22, 1 (2021), 63–75.

[5] Ahn, J., Nguyen, H., Campos, F. and Young, W. 2021. Transforming everyday information into practical analytics with crowdsourced assessment tasks. *LAK21: 11th International Learning Analytics and Knowledge Conference* (2021), 66–76.

[6] Alazaidah, R., Thabtah, F. and Al-Radaideh, Q. 2015. A multi-label classification approach based on correlations among labels. *International Journal of Advanced Computer Science and Applications*. 6, 2 (2015), 52–59.

[7] Aleven, V. and Koedinger, K.R. 2013. Knowledge component (KC) approaches to learner modeling. *Design recommendations for intelligent tutoring systems*. 1, (2013), 165–182.

[8] Almerico, G.M. and Baker, R.K. 2004. Bloom's Taxonomy illustrative verbs: Developing a comprehensive list for educator use. *Florida Association of Teacher Educators Journal*. 1, 4 (2004), 1–10.

[9] AlSumait, L., Barbará, D., Gentle, J. and Domeniconi, C. 2009. Topic significance ranking of LDA generative models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2009), 67–82.

[10] Amidei, J., Piwek, P. and Willis, A. 2018. Evaluation methodologies in Automatic Question Generation 2013-2018. *Proceedings of the 11th International Conference on Natural Language Generation* (2018), 307–317.

[11] Amidei, J., Piwek, P. and Willis, A. 2018. Rethinking the Agreement in Human Evaluation Tasks. *Proceedings of the 27th International Conference on Computational Linguistics* (2018), 3318–3329.

[12] Amini, N., Michoux, N., Warnier, L., Malcourant, E., Coche, E. and Berg, B.V. 2020. Inclusion of MCQs written by radiology residents in their annual evaluation: innovative method to enhance resident's empowerment? *Insights into Imaging*. 11, 1 (2020), 1–8.

[13] Andrade, A., Delandshere, G. and Danish, J.A. 2016. Using multimodal learning analytics to model student behavior: A systematic analysis of epistemological framing. *Journal of Learning Analytics*. 3, 2 (2016), 282–306.

[14] Arif, T., Asthana, S. and Collins-Thompson, K. 2024. Generation and Assessment of Multiple-Choice Questions from Video Transcripts using Large Language Models. *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta GA USA, Jul. 2024), 530–534.

[15] Armstrong, R.A. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*. 34, 5 (Sep. 2014), 502–508. DOI:https://doi.org/10.1111/opo.12131.

[16] Arora, S., Narayan, A., Chen, M.F., Orr, L., Guha, N., Bhatia, K., Chami, I., Sala, F. and Ré, C. 2022. Ask Me Anything: A simple strategy for prompting language models. arXiv.

[17] Assaly, I.R. and Smadi, O.M. 2015. Using Bloom's Taxonomy to Evaluate the

Cognitive Levels of Master Class Textbook's Questions. *English Language Teaching*. 8, 5 (2015), 100–110.

[18] Azevedo, J.M., Oliveira, E.P. and Beites, P.D. 2019. Using learning analytics to evaluate the quality of multiple-choice questions: A perspective with classical test theory and item response theory. *The International Journal of Information and Learning Technology*. 36, 4 (2019), 322–341.

[19] Badawy, M., Abd El-Aziz, A. and Hefny, H.A. 2016. Analysis of learning objectives for higher education textbooks using text mining. *2016 12th International Computer Engineering Conference (ICENCO)* (2016), 202–207.

[20] Bälter, O., Zimmaro, D. and Thille, C. 2018. Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments*. 5, 1 (2018), 1–19.

[21] Barnes, T. 2005. The q-matrix method: Mining student response data for knowledge. *American association for artificial intelligence 2005 educational data mining workshop* (2005), 1–8.

[22] Bates, S.P., Galloway, R.K., Riise, J. and Homer, D. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics-Physics Education Research*. 10, 2 (2014), 020105.

[23] Bathgate, M. and Schunn, C. 2017. The psychological characteristics of experiences that influence science motivation and content knowledge. *International Journal of Science Education*. 39, 17 (2017), 2402–2432.

[24] Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D. and Pritchard, D.E. 2012. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*. (2012).

[25] Bhowmick, A.K., Jagmohan, A., Vempaty, A., Dey, P., Hall, L., Hartman, J., Kokku, R. and Maheshwari, H. 2023. Automating Question Generation From Educational Text. *Artificial Intelligence XL*. M. Bramer and F. Stahl, eds. Springer Nature Switzerland. 437–450.

[26] Bier, N., Lip, S., Strader, R., Thille, C. and Zimmaro, D. 2014. An approach to knowledge component/skill modeling in online courses. *Open Learning*. (2014), 1–14.

[27] Bier, N., Moore, S. and Van Velsen, M. 2019. Instrumenting Courseware and Leveraging Data with the Open Learning Initiative. *Companion Proceedings 9th International Conference on Learning Analytics & Knowledge* (2019), 990–1001.

[28] Bier, N., Moore, S. and Van Velsen, M. 2019. Instrumenting courseware and leveraging data with the Open Learning Initiative (OLI). *Companion Proceedings 9th International Learning Analytics & Knowledge Conference, Tempe, AZ* (2019).

[29] Bird, S., Klein, E. and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

[30] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*. 3, Jan (2003), 993–1022.

[31] Breakall, J., Randles, C. and Tasker, R. 2019. Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*. 20, 2 (2019), 369–382.

[32] Brown, G.T. and Abdulnabi, H.H. 2017. Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education* (2017), 24.

[33] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., and others 2020. Language models are few-shot learners. *Advances in neural information processing systems*. 33, (2020), 1877–1901.

[34] Brunskill, E., Zimmaro, D. and Thille, C. 2018. Exploring the impact of the default

option on student engagement and performance in a statistics MOOC. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (2018), 1–4.

[35] Bulathwela, S., Muse, H. and Yilmaz, E. 2023. Scalable Educational Question Generation with Pre-trained Language Models. *Artificial Intelligence in Education*. N. Wang, G. Rebolledo-Mendez, N. Matsuda, O.C. Santos, and V. Dimitrova, eds. Springer Nature Switzerland. 327–339.

[36] Butler, A.C. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*. 7, 3 (2018), 323–331.

[37] Carvalho, P.F., Gao, M., Motz, B.A. and Koedinger, K.R. 2018. Analyzing the Relative Learning Benefits of Completing Required Activities and Optional Readings in Online Courses. *International Educational Data Mining Society*. (2018).

[38] Cen, H., Koedinger, K. and Junker, B. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. *International Conference on Intelligent Tutoring Systems* (2006), 164–175.

[39] Chan, A. 2022. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI and Ethics*. (2022), 1–12.

[40] Chau, H., Labutov, I., Thaker, K., He, D. and Brusilovsky, P. 2021. Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*. 31, 4 (2021), 820–846.

[41] Chen, G., Yang, J., Hauff, C. and Houben, G.-J. 2018. LearningQ: a large-scale dataset for educational question generation. *Twelfth International AAAI Conference on Web and Social Media* (2018).

[42] Chen, X., Breslow, L. and DeBoer, J. 2018. Analyzing productive learning behaviors for students using immediate corrective feedback in a blended learning environment. *Computers & Education*. 117, (2018), 59–74.

[43] Chin, C. and Brown, D.E. 2002. Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*. 24, 5 (2002), 521–549.

[44] Chipman, S.F., Schraagen, J.M. and Shalin, V.L. 2000. Introduction to cognitive task analysis. *Cognitive task analysis*. Psychology Press. 17–38.

[45] Clark, R. 2014. Cognitive task analysis for expert-based instruction in healthcare. *Handbook of research on educational communications and technology*. Springer. 541–551.

[46] Clifton, S.L. and Schriner, C.L. 2010. Assessing the quality of multiple-choice test items. *Nurse Educator*. 35, 1 (2010), 12–16.

[47] Cochran, K., Cohn, C., Hutchins, N., Biswas, G. and Hastings, P. 2022. Improving automated evaluation of formative assessments with text data augmentation. *International Conference on Artificial Intelligence in Education* (2022), 390–401.

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

[49] Costa, L.A., Salvador, L.N. and Amorim, R.R. 2018. Evaluation of academic performance based on learning analytics and ontology: a systematic mapping study. *2018 IEEE Frontiers in Education Conference (FIE)* (2018), 1–5.

[50] Costello, E., Holland, J. and Kirwan, C. 2018. The future of online testing and assessment: question quality in MOOCs. *International Journal of Educational Technology in Higher Education*. 15, 1 (Dec. 2018), 42. DOI:https://doi.org/10.1186/s41239-018-0124-z.

[51] Costello, E., Holland, J.C. and Kirwan, C. 2018. Evaluation of MCQs from MOOCs for common item writing flaws. *BMC Research Notes*. 11, 1 (Dec. 2018), 849. DOI:https://doi.org/10.1186/s13104-018-3959-4.

[52] Dai, P., Rzeszotarski, J.M., Paritosh, P. and Chi, E.H. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), 628–638.

[53] Danh, T., Desiderio, T., Herrmann, V., Lyons, H.M., Patrick, F., Wantuch, G.A. and Dell, K.A. 2020. Evaluating the quality of multiple-choice questions in a NAPLEX preparation book. *Currents in Pharmacy Teaching and Learning*. 12, 10 (2020), 1188–1193.

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

[55] Das, S., Mandal, S.K.D. and Basu, A. 2020. Identification of cognitive learning complexity of assessment questions using multi-class text classification. *Contemporary Educational Technology*. 12, 2 (2020), 275.

[56] DeCuir-Gunby, J.T., Marshall, P.L. and McCulloch, A.W. 2011. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field methods*. 23, 2 (2011), 136–155.

[57] Denny, P. 2015. Generating practice questions as a preparation strategy for introductory programming exams. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (2015), 278–283.

[58] Denny, P., Hamer, J., Luxton-Reilly, A. and Purchase, H. 2008. PeerWise: students sharing their multiple choice questions. *Proceedings of the Fourth international Workshop on Computing Education Research* (New York, NY, USA, Sep. 2008), 51–58.

[59] Denny, P., McDonald, F., Empson, R., Kelly, P. and Petersen, A. 2018. Empirical support for a causal relationship between gamification and learning outcomes. *Proceedings of the 2018 CHI conference on human factors in computing systems* (2018), 1–13.

[60] Denny, P., Sarsa, S., Hellas, A. and Leinonen, J. 2022. Robosourcing Educational Resources -- Leveraging Large Language Models for Learnersourcing. arXiv.

[61] Denny, P., Tempero, E., Garbett, D. and Petersen, A. 2017. Examining a student-generated question activity using random topic assignment. *Proceedings of the 2017 ACM conference on innovation and technology in computer science education* (2017), 146–151.

[62] Desmarais, M.C. and Naceur, R. 2013. A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. *International Conference on Artificial Intelligence in Education* (2013), 441–450.

[63] DiBattista, D. and Kurzawa, L. 2011. Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*. 2, 2 (2011), 4.

[64] Doroudi, S., Kamar, E. and Brunskill, E. 2019. Not everyone writes good examples but good examples can come from anywhere. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2019), 12–21.

[65] Doroudi, S., Kamar, E., Brunskill, E. and Horvitz, E. 2016. Toward a learning science for complex crowdsourcing tasks. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 2623–2634.

[66] Doughty, J. et al. 2024. A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education. *Proceedings of the 26th Australasian Computing Education Conference* (Sydney NSW Australia, Jan. 2024), 114–123.

[67] Douglas, B.D., Ewell, P.J. and Brauer, M. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*. 18, 3 (2023), e0279720.

[68] Douglas, M., Wilson, J. and Ennis, S. 2012. Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*. 49, 2 (2012), 111–121.

[69] Downing, S.M. 2005. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*. 10, 2 (2005), 133–143.

[70] Duret, D., Christley, R., Denny, P. and Senior, A. 2018. Collaborative learning with PeerWise. *Research in Learning Technology*. 26, (2018).

[71] Edens, K. and Potter, E. 2008. How students "unpack" the structure of a word problem: Graphic representations and problem solving. *School Science and Mathematics*. 108, 5 (2008), 184–196.

[72] Effenberger, T., Pelánek, R. and Čechák, J. 2020. Exploration of the robustness and generalizability of the additive factors model. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt Germany, Mar. 2020), 472–479.

[73] Egodawatte, G. 2010. A Rubric to Self-Assess and Peer-Assess Mathematical Problem Solving Tasks of College Students. *Acta didactica napocensia*. 3, 1 (2010), 75–88.

[74] Elkins, S., Kochmar, E., Cheung, J.C.K. and Serban, I. 2024. How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes. *Proceedings of the AAAI Conference on Artificial Intelligence* (2024).

[75] Elkins, S., Kochmar, E., Serban, I. and Cheung, J.C.K. 2023. How Useful Are Educational Questions Generated by Large Language Models? *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O.C. Santos, eds. Springer Nature Switzerland. 536–542.

[76] Elliott, V.F. 2018. Thinking about the coding process in qualitative data analysis. *Qualitative Report*. 23, 11 (2018).

[77] Farasat, A., Nikolaev, A., Miller, S. and Gopalsamy, R. 2017. Crowdlearning: Towards collaborative problem-posing at scale. *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (2017), 221–224.

[78] Flanagan, B., Tian, Z., Yamauchi, T., Dai, Y. and Ogata, H. 2024. A human-in-the-loop system for labeling knowledge components in Japanese mathematics exercises. *Research & Practice in Technology Enhanced Learning*. 19, (2024).

[79] Fletcher, S. and Islam, M.Z. 2018. Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*. 22, (2018).

[80] Ganda, D. and Buch, R. 2018. A survey on multi label classification. *Recent Trends in Programming Languages*. 5, 1 (2018), 19–23.

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

[82] Gledson, A., Apaolaza, A., Barthold, S., Günther, F., Yu, H. and Vigo, M. 2021. Characterising Student Engagement Modes through Low-Level Activity Patterns. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (2021), 88–97.

[83] Grainger, R., Osborne, E., Dai, W. and Kenwright, D. 2018. The process of developing a rubric to assess the cognitive complexity of student-generated multiple choice questions in medical education. *The Asia Pacific Scholar*. 3, 2 (2018), 19–24.

[84] Guilbert, J. 2002. The ambiguous and bewitching power of knowledge, skills and attitudes leads to confusing statements of learning objectives. *Education for*

Health. 15, 3 (2002), 362–369.

[85] Haladyna, T.M. 2004. *Developing and Validating Multiple-choice Test Items*. Psychology Press.

[86] Haladyna, T.M., Downing, S.M. and Rodriguez, M.C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*. 15, 3 (2002), 309–333.

[87] Hardy, J., Bates, S.P., Casey, M.M., Galloway, K.W., Galloway, R.K., Kay, A.E., Kirsop, P. and McQueen, H.A. 2014. Student-Generated Content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*. 36, 13 (Sep. 2014), 2180–2194. DOI:https://doi.org/10.1080/09500693.2014.916831.

[88] Haris, S.S. and Omar, N. 2012. A rule-based approach in Bloom's Taxonomy question classification through natural language processing. *2012 7th international conference on computing and convergence technology (ICCCT)* (2012), 410–414.

[89] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. Measuring Massive Multitask Language Understanding. *International Conference on Learning Representations*.

[90] Hodson, T.O. 2022. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*. 2022, (2022), 1–10.

[91] Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O.L. and Maritxalar, M. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. *Proceedings of The 12th Language Resources and Evaluation Conference* (2020), 1753–1762.

[92] Huang, A., Hancock, D., Clemson, M., Yeo, G., Harney, D., Denny, P. and Denyer, G. 2021. Selecting student-authored questions for summative assessments. *Research in Learning Technology*. 29, (2021).

[93] Huang, J., Zhang, Z., Qiu, J., Peng, L., Liu, D., Han, P. and Luo, K. 2021. Automatic Classroom Question Classification Based on Bloom's Taxonomy. *2021 13th International Conference on Education Technology and Computers* (2021), 33–39.

[94] Huang, Y., Aleven, V., McLaughlin, E. and Koedinger, K. 2020. A General Multi-method Approach to Design-Loop Adaptivity in Intelligent Tutoring Systems. *Artificial Intelligence in Education*. I.I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, eds. Springer International Publishing. 124–129.

[95] Hüllermeier, E., Fürnkranz, J., Loza Mencia, E., Nguyen, V.-L. and Rapp, M. 2020. Rule-based multi-label classification: Challenges and opportunities. *International Joint Conference on Rules and Reasoning* (2020), 3–19.

[96] Inglis, M., Palipana, A., Trenholm, S. and Ward, J. 2011. Individual differences in students' use of optional learning resources. *Journal of Computer Assisted Learning*. 27, 6 (2011), 490–502.

[97] Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y. and Yang, Y. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*. 36, (2024).

[98] Ji, T., Lyu, C., Jones, G., Zhou, L. and Graham, Y. 2022. QAScore—An Unsupervised Unreferenced Metric for the Question Generation Evaluation. *Entropy*. 24, 11 (2022), 1514.

[99] Jin, H., Chang, M. and Kim, J. 2019. SolveDeep: A System for Supporting Subgoal Learning in Online Math Problem Solving. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–6.

[100] Jones, J.A. 2019. Scaffolding self-regulated learning through student-generated quizzes. *Active Learning in Higher Education*. 20, 2 (2019), 115–126.

[101] Jury, B., Lorusso, A., Leinonen, J., Denny, P. and Luxton-Reilly, A. 2024. Evaluating LLM-generated Worked Examples in an Introductory Programming Course. *Proceedings of the 26th Australasian Computing Education Conference* (Sydney NSW Australia, Jan. 2024), 77–86.

[102] Kamalloo, E., Dziri, N., Clarke, C. and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023), 5591–5606.

[103] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S. and Hüllermeier, E. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*. 103, (2023), 102274.

[104] Kevian, D., Syed, U., Guo, X., Havens, A., Dullerud, G., Seiler, P., Qin, L. and Hu, B. 2024. Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra. arXiv.

[105] Khairani, A.Z. and Shamsuddin, H. 2016. Assessing Item Difficulty and Discrimination Indices of Teacher-Developed Multiple-Choice Tests. *Assessment for Learning Within and Beyond the Classroom*. Springer. 417–426.

[106] Khan, V.-J., Papangelis, K. and Markopoulos, P. 2020. Completing a Crowdsourcing Task Instead of an Assignment; What do University Students Think? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–8.

[107] Khosravi, H., Demartini, G., Sadiq, S. and Gasevic, D. 2021. Charting the design and analytics agenda of learnersourcing systems. *LAK21: 11th International Learning Analytics and Knowledge Conference* (2021), 32–42.

[108] Khosravi, H., Denny, P., Moore, S. and Stamper, J. 2023. Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*. (2023), 100151.

[109] Khosravi, H., Kitto, K. and Williams, J.J. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics*. 6, 3 (2019), 91–105.

[110] Kim, J. 2015. *Learnersourcing: improving learning with collective learner activity*. Massachusetts Institute of Technology.

[111] Kim, M.K. 2015. Models of learning progress in solving complex problems: Expertise development in teaching and learning. *Contemporary Educational Psychology*. 42, (2015), 1–16.

[112] Kim, M.-K., Patel, R.A., Uchizono, J.A. and Beck, L. 2012. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American journal of pharmaceutical education*. 76, 6 (2012).

[113] Kittur, A., Chi, E.H. and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the SIGCHI conference on human factors in computing systems* (2008), 453–456.

[114] Kizilcec, R.F. and Halawa, S. 2015. Attrition and achievement gaps in online learning. *Proceedings of the second (2015) ACM conference on learning@ scale* (2015), 57–66.

[115] Koedinger, K. and McLaughlin, E. 2010. Seeing language learning inside the math: Cognitive analysis yields transfer. *Proceedings of the annual meeting of the cognitive science society* (2010).

[116] Koedinger, K.R., Corbett, A.T. and Perfetti, C. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*. 36, 5 (2012), 757–798.

[117] Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A. and Bier, N.L. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC.

Proceedings of the second (2015) ACM conference on learning@ scale (2015), 111–120.

[118] Koedinger, K.R., McLaughlin, E.A. and Stamper, J.C. 2012. Automated Student Model Improvement. *International Educational Data Mining Society*. (2012).

[119] Krathwohl, D.R. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*. 41, 4 (Nov. 2002), 212–218. DOI:https://doi.org/10.1207/s15430421tip4104_2.

[120] Kremer, I., Mansour, Y. and Perry, M. 2014. Implementing the "wisdom of the crowd." *Journal of Political Economy*. 122, 5 (2014), 988–1012.

[121] Krishna, K., Wieting, J. and Iyyer, M. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), 737–762.

[122] Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*. 30, 1 (Mar. 2020), 121–204. DOI:https://doi.org/10.1007/s40593-019-00186-y.

[123] Kurtz, J.B., Lourie, M.A., Holman, E.E., Grob, K.L. and Monrad, S.U. 2019. Creating assessments as an active learning strategy: what are students' perceptions? A mixed methods study. *Medical education online*. 24, 1 (2019), 1630239.

[124] Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *biometrics*. (1977), 159–174.

[125] van der Lee, C., Gatt, A., van Miltenburg, E. and Krahmer, E. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*. 67, (May 2021), 101151. DOI:https://doi.org/10.1016/j.csl.2020.101151.

[126] Lee, D.D. and Seung, H.S. 2001. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* (2001), 556–562.

[127] Lee, P., Bubeck, S. and Petro, J. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*. 388, 13 (2023), 1233–1239.

[128] Lee, R.L. 2004. *The impact of cognitive task analysis on performance: A meta-analysis of comparative studies*. University of Southern California.

[129] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, W.B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016), 110–119.

[130] Li, Z., Pardos, Z.A. and Ren, C. 2024. Aligning open educational resources to new taxonomies: How AI technologies can help and in which scenarios. *Computers & Education*. 216, (2024), 105027.

[131] Li, Z., Ren, C., Li, X. and Pardos, Z.A. 2021. Learning Skill Equivalencies Across Platform Taxonomies. *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine CA USA, Apr. 2021), 354–363.

[132] Lipton, Z.C., Elkan, C. and Narayanaswamy, B. 2014. Thresholding classifiers to maximize F1 score. *stat*. 1050, (2014), 14.

[133] Liu, R. and Koedinger, K.R. 2017. Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains. *Journal of Educational Data Mining*. 9, 1 (2017), 25–41.

[134] Liu, R., McLaughlin, E.A. and Koedinger, K.R. 2014. Interpreting model discovery and testing generalization to a new dataset. *Educational Data Mining 2014* (2014).

[135] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C. 2023. GPTEval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634*. (2023).

[136] Liu, Z., Chen, C., Wang, J., Chen, M., Wu, B., Che, X., Wang, D. and Wang, Q. 2024.

Make LLM a Testing Expert: Bringing Human-like Interaction to Mobile GUI Testing via Functionality-aware Decisions. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (Lisbon Portugal, Apr. 2024), 1–13.

[137] Long, J. 2023. Large Language Model Guided Tree-of-Thought. arXiv.

[138] Long, Y. and Aleven, V. 2013. Supporting students' self-regulated learning with an open learner model in a linear equation tutor. *International conference on artificial intelligence in education* (2013), 219–228.

[139] Long, Y., Holstein, K. and Aleven, V. 2018. What exactly do students learn when they practice equation solving?: refining knowledge components with the additive factors model. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (Sydney New South Wales Australia, Mar. 2018), 399–408.

[140] Lu, O.H., Huang, A.Y., Tsai, D.C. and Yang, S.J. 2021. Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students Learning Performance. *Educational Technology & Society*. 24, 3 (2021), 159–173.

[141] Lu, X., Fan, S., Houghton, J., Wang, L. and Wang, X. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany, Apr. 2023), 1–18.

[142] Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A. and Zheng, Q. 2017. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. *Thirty-first AAAI conference on artificial intelligence* (2017).

[143] Markel, J.M., Opferman, S.G., Landay, J.A. and Piech, C. 2023. GPTeach: Interactive TA Training with GPT-based Students. *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen Denmark, Jul. 2023), 226–236.

[144] Mathur, N., Baldwin, T. and Cohn, T. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 4984–4997.

[145] Matsuda, N., Furukawa, T., Bier, N. and Faloutsos, C. 2015. Machine Beats Experts: Automatic Discovery of Skill Models for Data-Driven Online Course Refinement. *International Educational Data Mining Society*. (2015).

[146] Matsuda, N., Wood, J., Shrivastava, R., Shimmei, M. and Bier, N. 2022. Latent skill mining and labeling from courseware content. *Journal of educational data mining*. 14, 2 (2022).

[147] Mavis, B.E., Cole, B.L. and Hoppe, R.B. 2001. A survey of student assessment in US medical schools: the balance of breadth versus fidelity. *Teaching and Learning in Medicine*. 13, 2 (2001), 74–79.

[148] McHugh, M.L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*. 22, 3 (2012), 276–282.

[149] McQueen, H.A., Shields, C., Finnegan, D., Higham, J. and Simmen, M. 2014. PeerWise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention. *Biochemistry and Molecular Biology Education*. 42, 5 (2014), 371–381.

[150] Mitros, P. 2015. Learnersourcing of complex assessments. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (2015), 317–320.

[151] Moeen-uz-Zafar Khan, B.M. 2011. Evaluation of modified essay questions (MEQ) and multiple choice questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. *International journal of health sciences*. 5, 1 (2011), 39.

[152] Monrad, S.U., Bibler Zaidi, N.L., Grob, K.L., Kurtz, J.B., Tai, A.W., Hortsch, M., Gruppen, L.D. and Santen, S.A. 2021. What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. *Medical*

Teacher. 43, 5 (May 2021), 575–582. DOI:https://doi.org/10.1080/0142159X.2021.1879376.

[153] Moon, H., Yang, Y., Yu, H., Lee, S., Jeong, M., Park, J., Shin, J., Kim, M. and Choi, S. 2022. Evaluating the Knowledge Dependency of Questions. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), 10512–10526.

[154] Moore, S., Costello, E., Nguyen, H.A. and Stamper, J. 2024. An Automatic Question Usability Evaluation Toolkit. *Artificial Intelligence in Education* (Cham, 2024), 31–46.

[155] Moore, S., Fang, E., Nguyen, H.A. and Stamper, J. 2023. Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom's Taxonomy. *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen Denmark, Jul. 2023), 25–34.

[156] Moore, S., Nguyen, H. and Stamper, J. 2023. Students' Domain Confidence and their Participation in Optional Learnersourcing Activities. *Companion Proceedings in LAK23: 13th International Learning Analytics and Knowledge Conference* (Arlington, TX, 2023).

[157] Moore, S., Nguyen, H. and Stamper, J. Utilizing Crowdsourcing and Topic Modeling to Generate Knowledge Components for Math and Writing Problems. *Proceedings of the 28th International Conference on Computers in Education* 31–40.

[158] Moore, S., Nguyen, H.A., Bier, N., Domadia, T. and Stamper, J. 2022. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-Tel 2022, Toulouse, France, September 12–16, 2022, Proceedings* (2022), 243–257.

[159] Moore, S., Nguyen, H.A., Bier, N., Domadia, T. and Stamper, J. 2023. Who writes tomorrow's learning activities? Exploring community college student participation in learnersourcing. *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023, pp. 664-671* (2023).

[160] Moore, S., Nguyen, H.A., Chen, T. and Stamper, J. 2023. Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. *Responsive and Sustainable Educational Futures*. O. Viberg, I. Jivet, P.J. Muñoz-Merino, M. Perifanou, and T. Papathoma, eds. Springer Nature Switzerland. 229–245.

[161] Moore, S., Nguyen, H.A. and Stamper, J. 2020. Evaluating Crowdsourcing and Topic Modeling in Generating Knowledge Components from Explanations. *International Conference on Artificial Intelligence in Education* (2020), 398–410.

[162] Moore, S., Nguyen, H.A. and Stamper, J. 2021. Examining the Effects of Student Participation and Performance on the Quality of Learnersourcing Multiple-Choice Questions. *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event Germany, Jun. 2021), 209–220.

[163] Moore, S., Nguyen, H.A. and Stamper, J. 2022. Leveraging Students to Generate Skill Tags that Inform Learning Analytics. *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 791-798* (2022).

[164] Moore, S., Nguyen, H.A. and Stamper, J. 2020. Towards Crowdsourcing the Identification of Knowledge Components. *Proceedings of the Seventh ACM Conference on Learning @ Scale* (2020), 245–248.

[165] Moore, S., Norman, B. and Stamper, J. Assessing Educational Quality: Comparative Analysis of Crowdsourced, Expert, and AI-Driven Rubric Applications. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

[166] Moore, S., Schmucker, R., Mitchell, T. and Stamper, J. 2024. Automated Generation and Tagging of Knowledge Components from Multiple-Choice Questions. *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (New York, NY, USA, Jul. 2024), 122–133.

[167] Moore, S. and Stamper, J. 2019. Decision support for an adversarial game environment using automatic hint generation. *International Conference on Intelligent Tutoring Systems* (2019), 82–88.

[168] Moore, S., Stamper, J., Brooks, C., Denny, P. and Khosravi, H. 2022. Learnersourcing: Student-generated Content @ Scale. *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City NY USA, Jun. 2022), 259–262.

[169] Mooris, J. 2022. Python Language Tool.

[170] Mulla, N. and Gharpure, P. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*. 12, 1 (Mar. 2023), 1–32. DOI:https://doi.org/10.1007/s13748-023-00295-9.

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

[172] Ng, E.M. 2014. Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring. *Computers & education*. 73, (2014), 141–148.

[173] Nguyen, H., Wang, Y., Stamper, J. and McLaren, B.M. 2019. Using Knowledge Component Modeling to Increase Domain Understanding in a Digital Learning Game. *Proceedings of the 12th International Conference on Educational Data Mining* (2019), 139–148.

[174] Ni, L., Bao, Q., Li, X., Qi, Q., Denny, P., Warren, J., Witbrock, M. and Liu, J. 2022. Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), 12826–12834.

[175] OpenAI 2023. GPT-4 Technical Report. arXiv.

[176] PaaBen, B., Dywel, M., Fleckenstein, M. and Pinkwart, N. 2022. Sparse Factor Autoencoders for Item Response Theory. *International Educational Data Mining Society*. (2022).

[177] Palan, S. and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*. 17, (2018), 22–27.

[178] Paolacci, G., Chandler, J. and Ipeirotis, P.G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*. 5, 5 (2010), 411–419.

[179] Pardos, Z., Bergner, Y., Seaton, D. and Pritchard, D. 2013. Adapting bayesian knowledge tracing to a massive open online course in edx. *Educational Data Mining 2013* (2013).

[180] Pardos, Z.A. and Dadu, A. 2017. Imputing KCs with representations of problem content and context. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (2017), 148–155.

[181] Pate, A. and Caldwell, D.J. 2014. Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning*. 6, 1 (2014), 130–134.

[182] Patikorn, T., Deisadze, D., Grande, L., Yu, Z. and Heffernan, N. 2019. Generalizability of Methods for Imputing Mathematical Skills Needed to Solve Problems from Texts. *International Conference on Artificial Intelligence in Education* (2019), 396–405.

[183] Paulin, D. and Haythornthwaite, C. 2016. Crowdsourcing the curriculum: Redefining e-learning practices through peer-generated approaches. *The Information Society*. 32, 2 (2016), 130–142.

[184] Pham, H., Besanko, J. and Devitt, P. 2018. Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. *MedEdPublish*. 7, (2018), 225.

[185] Pi, S., An, X., Xu, S. and Li, J. 2020. A Comparative Study on Three Multi-Label Classification Tools. *Proceedings of the 2020 3rd International Conference on Information Management and Management Science* (London United Kingdom, Aug. 2020), 8–12.

[186] Pugh, D., De Champlain, A., Gierl, M., Lai, H. and Touchie, C. 2020. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning*. 15, 1 (2020), 1–13.

[187] Qasrawi, R. and BeniAbdelrahman, A. 2020. The Higher and Lower-Order Thinking Skills (HOTS and LOTS) in Unlock English Textbooks (1st and 2nd Editions) Based on Bloom’s Taxonomy: An Analysis Study. *International Online Journal of Education and Teaching*. 7, 3 (2020), 744–758.

[188] Rahman, S.A. and Manaf, N.F.A. 2017. A Critical Analysis of Bloom’s Taxonomy in Teaching Creative and Critical Thinking Skills in Malaysia through English Literature. *English Language Teaching*. 10, 9 (2017), 245–256.

[189] Raina, V. and Gales, M. 2022. Multiple-Choice Question Generation: Towards an Automated Assessment Framework. arXiv.

[190] Ratnayake, H. and Wang, C. 2024. A Prompting Framework to Enhance Language Model Output. *AI 2023: Advances in Artificial Intelligence*. T. Liu, G. Webb, L. Yue, and D. Wang, eds. Springer Nature Singapore. 66–81.

[191] Rau, M.A. 2017. Do Knowledge-Component Models Need to Incorporate Representational Competencies? *International Journal of Artificial Intelligence in Education*. 27, 2 (Jun. 2017), 298–319. DOI:https://doi.org/10.1007/s40593-016-0134-8.

[192] Riggs, C.D., Kang, S. and Rennie, O. 2020. Positive Impact of Multiple-Choice Question Authoring and Regular Quiz Participation on Student Learning. *CBE—Life Sciences Education*. 19, 2 (2020), ar16.

[193] Rizvi, S., Rienties, B. and Khoja, S.A. 2019. The role of demographics in online learning; A decision tree based approach. *Computers & Education*. 137, (2019), 32–47.

[194] Rowan, B. and White, M. 2022. The Common Core State Standards Initiative as an Innovation Network. *American Educational Research Journal*. 59, 1 (Feb. 2022), 73–111. DOI:https://doi.org/10.3102/00028312211006689.

[195] Ruseti, S., Dascalu, M., Johnson, A.M., Balyan, R., Kopp, K.J., McNamara, D.S., Crossley, S.A. and Trausan-Matu, S. 2018. Predicting question quality using recurrent neural networks. *International conference on artificial intelligence in education* (2018), 491–502.

[196] Rush, B.R., Rankin, D.C. and White, B.J. 2016. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*. 16, 1 (2016), 1–10.

[197] Ruthotto, I., Kreth, Q., Stevens, J., Trively, C. and Melkers, J. 2020. Lurking and participation in the virtual classroom: The effects of gender, race, and age among graduate students in computer science. *Computers & Education*. 151, (2020), 103854.

[198] Ryan, S., Kaufman, J., Greenhouse, J., She, R. and Shi, J. 2016. The effectiveness of blended online learning courses at the community college level. *Community College Journal of Research and Practice*. 40, 4 (2016), 285–298.

[199] Sachdeva, A.K. 1996. Use of effective questioning to enhance the cognitive abilities of students. *Journal of Cancer Education*. 11, 1 (1996), 17–24.

[200] Safranek, C.W., Sidamon-Eristoff, A.E., Gilson, A. and Chartash, D. 2023. The role of large language models in medical education: applications and implications. *JMIR Medical Education*. JMIR Publications Toronto, Canada.

[201] Saldana, J. 2009. An introduction to codes and coding. *The coding manual for*

qualitative researchers. 3, (2009).

[202] Sarsa, S., Denny, P., Hellas, A. and Leinonen, J. 2022. Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models. *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1* (2022), 27–43.

[203] Sauerwein, A.M. and Wegner, J.R. 2020. Using think-alouds to uncover expert-novice gaps in AAC intervention planning skills. *Teaching and Learning in Communication Sciences & Disorders*. 4, 2 (2020), 8.

[204] Schmidt-Weigand, F., Kohnert, A. and Glowalla, U. 2010. A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and instruction*. 20, 2 (2010), 100–110.

[205] Schurmeier, K.D., Atwood, C.H., Shepler, C.G. and Lautenschlager, G.J. 2010. Using item response theory to assess changes in student performance based on changes in question wording. *Journal of chemical education*. 87, 11 (2010), 1268–1272.

[206] Scialom, T. and Staiano, J. 2020. Ask to Learn: A Study on Curiosity-driven Question Generation. *Proceedings of the 28th International Conference on Computational Linguistics* (2020), 2224–2235.

[207] Scully, D. 2019. Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*. 22, 1 (2019), 4.

[208] Shabana, K.M. and Lakshminarayanan, C. 2023. Unsupervised Concept Tagging of Mathematical Questions from Student Explanations. *Artificial Intelligence in Education*. N. Wang, G. Rebolledo-Mendez, N. Matsuda, O.C. Santos, and V. Dimitrova, eds. Springer Nature Switzerland. 627–638.

[209] Shaikh, S., Daudpotta, S.M. and Imran, A.S. 2021. Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings. *IEEE Access*. 9, (2021), 117887–117909.

[210] Shen, J.T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S. and Lee, D. 2021. Classifying Math Knowledge Components via Task-Adaptive Pre-Trained BERT. *International Conference on Artificial Intelligence in Education* (2021), 408–419.

[211] Shi, Y., Schmucker, R., Chi, M., Barnes, T. and Price, T. 2023. KC-Finder: Automated Knowledge Component Discovery for Programming Problems. *International Educational Data Mining Society*. (2023).

[212] Singh, A., Brooks, C. and Doroudi, S. 2022. Learnersourcing in Theory and Practice: Synthesizing the Literature and Charting the Future. *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City NY USA, Jun. 2022), 234–245.

[213] Singh, A., Brooks, C., Lin, Y. and Li, W. 2021. What's In It for the Learners? Evidence from a Randomized Field Experiment on Learnersourcing Questions in a MOOC. *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event Germany, Jun. 2021), 221–233.

[214] Singh, A., Brooks, C. and Wang, X. 2024. The Impact of Student-AI Collaborative Feedback Generation on Learning Outcomes. *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI* (2024).

[215] Singh, A., Brooks, C., Wang, X., Li, W., Kim, J. and Wilson, D. 2024. Bridging Learnersourcing and AI: Exploring the Dynamics of Student-AI Collaborative Feedback Generation. *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto Japan, Mar. 2024), 742–748.

[216] Stamper, J.C. and Koedinger, K.R. 2011. Human-machine student model discovery and improvement using DataShop. *International Conference on Artificial Intelligence in Education* (2011), 353–360.

[217] Steuer, T., Bongard, L., Uhlig, J. and Zimmer, G. 2021. On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine

Translation. *European Conference on Technology Enhanced Learning* (2021), 289–294.

[218] Straková, J., Straka, M. and Hajic, J. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2014), 13–18.

[219] Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J. and Lin, J. 2024. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. arXiv.

[220] Sun, L., Liu, Y., Joseph, G., Yu, Z., Zhu, H. and Dow, S.P. 2022. Comparing experts and novices for ai data work: Insights on allocating human intelligence to design a conversational agent. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2022), 195–206.

[221] Tarrant, M., Knierim, A., Hayes, S.K. and Ware, J. 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*. 26, 8 (2006), 662–671.

[222] Tarrant, M. and Ware, J. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education*. 42, 2 (2008), 198–206.

[223] Thiergart, J., Huber, S. and Übellacker, T. 2021. Understanding Emails and Drafting Responses–An Approach Using GPT-3. *arXiv e-prints*. (2021), arXiv-2102.

[224] Tian, Z., Flanagan, B., Dai, Y. and Ogata, H. 2022. Automated matching of exercises with knowledge components. *30th International Conference on Computers in Education Conference Proceedings* (2022), 24–32.

[225] Tsoumakas, G. and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*. 3, 3 (2007), 1–13.

[226] Wang, M., Chau, H., Thaker, K., Brusilovsky, P. and He, D. 2021. Knowledge Annotation for Intelligent Textbooks. *Technology, Knowledge and Learning*. (2021), 1–22.

[227] Wang, X., Talluri, S.T., Rose, C. and Koedinger, K. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago IL USA, Jun. 2019), 1–10.

[228] Wang, Z., Funakoshi, K. and Okumura, M. 2023. Automatic Answerability Evaluation for Question Generation. arXiv.

[229] Wang, Z., Manning, K., Mallick, D.B. and Baraniuk, R.G. 2021. Towards Blooms Taxonomy Classification Without Labels. *International Conference on Artificial Intelligence in Education* (2021), 433–445.

[230] Wang, Z., Valdez, J., Basu Mallick, D. and Baraniuk, R.G. 2022. Towards Human-Like Educational Question Generation with Large Language Models. *Artificial Intelligence in Education*. M.M. Rodrigo, N. Matsuda, A.I. Cristea, and V. Dimitrova, eds. Springer International Publishing. 153–166.

[231] Wang, Z., Zhang, W., Liu, N. and Wang, J. 2021. Scalable rule-based representation learning for interpretable classification. *Advances in Neural Information Processing Systems*. 34, (2021), 30479–30491.

[232] Weld, D.S., Adar, E., Chilton, L., Hoffmann, R., Horvitz, E., Koch, M., Landay, J., Lin, C.H. and Mausam, M. 2012. Personalized online education—a crowdsourcing challenge. *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

[233] Williams, J.J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K.Z., Lasecki, W.S. and Heffernan, N. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (2016), 379–388.

[234] Xie, J., Peng, N., Cai, Y., Wang, T. and Huang, Q. 2021. Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 30, (2021), 280–291.

[235] Yahya, A.A., Toukal, Z. and Osman, A. 2012. Bloom's Taxonomy–based classification for item bank questions using support vector machines. *Modern advances in intelligent systems and tools*. Springer. 135–140.

[236] Yavuz, F., Çelik, Ö. and Yavaş Çelik, G. 2024. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*. (Jun. 2024), bjet.13494. DOI:https://doi.org/10.1111/bjet.13494.

[237] Yeckehzaare, I., Barghi, T. and Resnick, P. 2020. QMaps: Engaging Students in Voluntary Question Generation and Linking. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[238] Yu, F.Y. and Cheng, W.W. 2020. Effects of academic achievement and group composition on quality of student-generated questions and use patterns of online procedural prompts. *28th International Conference on Computers in Education, ICCE 2020* (2020), 573–581.

[239] Yu, F.-Y. and Liu, Y.-H. 2009. Creating a psychologically safe online space for a student-generated questions learning activity via different identity revelation modes. *British Journal of Educational Technology*. 40, 6 (2009), 1109–1123.

[240] Zaidi, N.L.B., Grob, K.L., Monrad, S.M., Kurtz, J.B., Tai, A., Ahmed, A.Z., Gruppen, L.D. and Santen, S.A. 2018. Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work? *Academic Medicine*. 93, 6 (2018), 856–859.

[241] Zhang, D., Wang, J. and Zhao, X. 2015. Estimating the Uncertainty of Average F1 Scores. *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton Massachusetts USA, Sep. 2015), 317–320.

[242] Zhang, J., Wong, C., Giacaman, N. and Luxton-Reilly, A. 2021. Automated classification of computing education questions using Bloom's taxonomy. *Australasian Computing Education Conference* (2021), 58–65.

[243] Zheng, A.Y., Lawhorn, J.K., Lumley, T. and Freeman, S. 2008. Application of Bloom's Taxonomy Debunks the"MCAT Myth." *SCIENCE-NEW YORK THEN WASHINGTON-*. 319, 5862 (2008), 414.