Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom's Taxonomy

Steven Moore
Human-Computer Interaction
Carnegie Mellon University
Pittsburgh, PA, USA
StevenJamesMoore@gmail.com

Ellen Fang
Human-Computer Interaction
Carnegie Mellon University
Pittsburgh, PA, USA
ellenf@andrew.cmu.edu

Huy A. Nguyen
Human-Computer Interaction
Carnegie Mellon University
Pittsburgh, PA, USA
hn1@andrew.cmu.edu

John Stamper
Human-Computer Interaction
Carnegie Mellon University
Pittsburgh, PA, USA
john@stamper.org

ABSTRACT

Multiple-choice questions, which are widely used in educational assessments, have the potential to negatively impact student learning and skew analytics when they contain item-writing flaws. Existing methods for evaluating multiple-choice questions in educational contexts tend to focus primarily on machine readability metrics, such as grammar, syntax, and formatting, without considering the intended use of the questions within course materials and their pedagogical implications. In this study, we present the results of crowdsourcing the evaluation of multiple-choice questions based on 15 common item-writing flaws. Through analysis of 80 crowdsourced evaluations on questions from the domains of calculus and chemistry, we found that crowdworkers were able to accurately evaluate the questions, matching 75% of the expert evaluations across multiple questions. They were able to correctly distinguish between two levels of Bloom's Taxonomy for the calculus questions, but were less accurate for chemistry questions. We discuss how to scale this question evaluation process and the implications it has across other domains. This work demonstrates how crowdworkers can be leveraged in the quality evaluation of educational questions, regardless of prior experience or domain knowledge.

CCS CONCEPTS

•Information systems~Crowdsourcing • Applied computing ~Education • Human-centered computing~Collaborative content creation

L@S '23, July 20-22, 2023, Copenhagen, Denmark
© 2023 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0025-5/23/07...$15.00
https://doi.org/10.1145/3573051.3593396

KEYWORDS

crowdsourcing, learnersourcing, question evaluation, question quality, question generation

ACM Reference format:

Steven Moore, Ellen Fang, Huy A. Nguyen and John Stamper. 2023. Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom's Taxonomy. In *Proceedings of ACM Learning @ Scale conference (L@S'23). ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/3573051.3593396

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 [11]. 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 [10]. However, a continually growing research effort has led to the advancement of MCQ authoring methods that do not rely on experts [27]. 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 [36]. 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 [32, 47, 56].

These popular methods allow for the scaling of educational MCQ creation, but they are highly susceptible to generating questions that contain detrimental flaws [3, 29]. 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, 16]. 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 [6]. 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 [55]. 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 [12]. 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 [20].

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 [42, 45]. 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 [9, 60]. 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 [43, 49].

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 Taxonomy [34]. Using the wisdom of the crowds, we evaluated if the majority response aligned with expert evaluation of the same questions. This study investigates the following research questions: (1) *How effective are crowdworkers at applying the IWF rubric to assess multiple-choice questions?* (2) *To what extent can crowdworkers accurately identify the cognitive level of multiple-choice 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.

2 Related Work

We first highlight the significance of Bloom's Revised Taxonomy with MCQs. Then we review previous work on data-driven methods for evaluating the quality of multiple-choice questions. Next, we discuss common methods of human evaluation that the data-driven methods often get compared against. Finally, we review the existing literature around crowdsourcing techniques in education and the advantages these methods can bring to scaling this process.

2.1 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 [28]. 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 [34]. Previous research has shown that MCQs commonly assess lower-level cognitive processes, such as recall, but they can assess all levels [17]. Assigning a Bloom's Revised Taxonomy label to each question can improve problem selection and learning analytics [13]. Automated methods for determining the cognitive level of questions have shown promise, with accuracy as high as 84% compared to human labels [41]. However, these methods often require large amounts of training data or expert time, making them inaccessible and difficult to scale.

2.2 Data-driven Evaluation of Questions

All multiple-choice questions, regardless of whether they are created by instructors, students, or through automation, are vulnerable to defects that affect their effectiveness and reliability [53]. Evaluating the quality of MCQs can be subjective, as there is no clear consensus on the criteria that make a question high-quality and appropriate for educational use. To address this issue, researchers have utilized item response theory and statistical methods to evaluate the quality and pedagogical usefulness of MCQs [30, 36]. However, these techniques rely on post-hoc analysis of student performance data, which can hinder the learning process if the questions have not been properly evaluated for quality beforehand [12, 46]. A question's difficulty might not be attributed to the content it assesses, but rather that the question was poorly formulated, causing the student to struggle. For instance, [54] demonstrated that questions containing text with grammatical errors or a wording that the students find confusing can cause students to select an incorrect response, even when they have demonstrated mastery of the content. This can lead to students wasting time answering poor-quality questions that may have a negative impact on their performance and achievement within a course. To ensure student learning is not hindered and to make effective use of student time, it is important to evaluate MCQs before they are used [50].

As advancements in machine learning and natural language processing continue, the accuracy of automatic question evaluation methods approaches human parity for certain domains

[36]. The automatic evaluation methods often utilize metrics related to readability and explainability, using metrics such as BLUE or METEOR [55]. Previous research has demonstrated how these metrics often do not correlate with expert human evaluation [38]. These surface-level metrics neglect the pedagogical implications of the questions, such as how students might answer the question based on the wording, answer choices, or cues that can be used to guess the correct answer. Additionally, a plethora of the automatic question evaluation work utilizes simplistic multiple-choice questions from lower grade-levels and lower cognitive levels, such as recall [18, 37]. The accuracy of these methods often heavily decreases when questions are used from more complex domains, such as questions commonly found in higher education STEM courses [8]. However, the standard against which these automatic evaluation methods are measured is human evaluation [21].

2.3 Human Evaluation of Questions

The use of human evaluation for multiple-choice questions consists of having one or more experts assess the question based on a set of criteria [39]. The criteria used for evaluation can vary greatly depending on the study, ranging from subjective assessments by experts, such as whether they would use the question in their classroom, to more standardized methods, such as using a rubric [9, 25]. Several popular rubrics have been used for the evaluation process, but the item-writing flaws (IWF) rubric, which contains multiple criteria for the evaluation of educational questions, has been standardized and validated through previous research [53, 59]. In a previous study by [9], the 15-item IWF rubric was applied by expert evaluators to over 1,000 MCQs from standardized chemistry exams used at the undergraduate level. They found that 83% of the MCQs contained at least one IWF, with the most common flaw being the inclusion of implausible distractors. Although the IWF rubric is effective for evaluating educational questions, its application often involves significant human effort and can be time-consuming, especially when evaluating many questions across various subject areas. Additionally, previous studies report only utilizing evaluators with domain-expertise to apply the IWF rubric. In the present study, we utilize the IWF rubric, but enlist crowdworkers to apply it to educational questions rather than instructor or domain experts.

Learnersourcing approaches often involve students in the creation and evaluation of MCQs across a variety of domains [19, 40]. This offers a scalable solution for both processes, while also being a beneficial learning experience for the students [57]. However, the criteria students use for this evaluation is often ill defined and the research around the evaluation process being beneficial to student learning is sparse, as often the generation of the MCQ is what contributes to student learning [3]. Additionally, previous work remains unclear how the students' background, prior knowledge, and other external factors influence their evaluation of other student-generated MCQs. While these methods have gaps that make it unclear how beneficial the evaluation process is, these methods do demonstrate how with proper scaffolding or task design, individuals that do not necessarily have domain expertise can effectively be leveraged in the evaluation of educational MCQs.

2.4 Crowdsourcing in Education

Crowdsourcing has been used in educational contexts to scale the generation of questions, provide feedback and hints, make improvements to the question content, and label the prior knowledge needed to correctly answer a question [1, 24, 31, 44]. A common challenge in educational crowdsourcing is how to effectively increase the quality of the work, while maintaining the scalability [61]. Crowdsourcing tasks often require specialized knowledge, such as domain expertise or knowledge of pedagogical approaches, that a crowdworker might not possess [33]. To help overcome this, previous research has provided crowdworkers with expert examples and defined rubrics that they can leverage during their crowdsourcing task [22]. This has been shown to increase the quality of the crowdworker's response to the task, reduce the time spent, and help alleviate some of the difficulty that may arise from a lack of domain or task expertise [23]. For instance, [4] provided crowdworkers that lacked background knowledge about the teaching and learning of writing with a rubric used to assess students' writing artifacts. They found that their crowdsourced scores of the writing artifacts demonstrated substantial agreement with expert evaluation. The utilization of rubrics in crowdsourcing tasks serves two purposes: it establishes standardized evaluation criteria and effectively harnesses the diverse backgrounds and multiple perspectives of the crowd. In the current study, a rubric is used to provide guidance for crowdworkers who may not possess the domain knowledge typically required to assess the quality of educational questions.

3 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 [48]. 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 [14].

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.

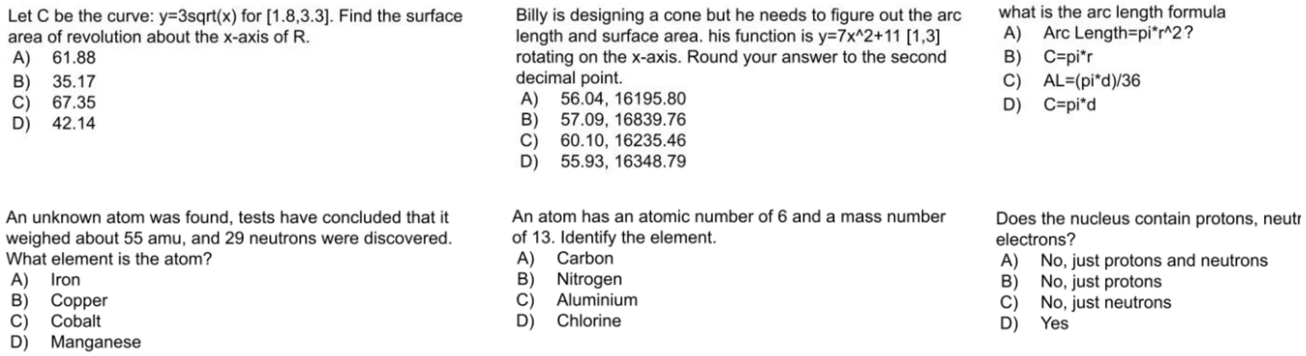
3.1 Calculus & Chemistry Questions

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 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.

3.2 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 [26]. 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 [9, 15, 49]. A complete list of the 15 criteria that make up the rubric can be found in Table 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 [8, 28]. 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.

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



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, neutr electrons?
A) No, just protons and neutrons
B) No, just protons
C) No, just neutrons
D) Yes

Figure 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.

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 [7] 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.

3.3 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

Table 1: The rubric of 15 item-writing flaws used to evaluate the multiple-choice questions.

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

it is from the majority. This is known as the *wisdom of the crowd* and is a popular method used to aggregate crowdsourced responses [35].

4 Results

4.1 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 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*.

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 cue*s.

Table 2: The percentage of crowdworkers that evaluated each IWF criteria the same as the expert evaluators for the given question.

	Chemistry			Calculus		
Criteria	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

4.2 Cognitive Level Accuracy

The average number of questions the crowdworkers correctly identified the cognitive levels of can be seen in Table 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.*

Table 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.

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%

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.

5 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 [58]. 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 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 [62]. A previous study by [59] 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 [42]. Additionally, the agreement threshold of 50% or higher with the expert evaluation aligned with prior crowdsourcing research [35]. 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 [5], 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 [8]. 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 [52]. 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 Taxonomy [51].

The chemistry question that crowdworkers misidentified the cognitive level of can be found in the bottom left of Figure 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.

6 Limitations and 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 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 [28], 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 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.

7 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.

REFERENCES

[1] Abdi, S., Khosravi, H. and Sadiq, S. 2020. Modelling 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] Aflalo, E. 2021. Students generating questions as a way of learning. *Active Learning in Higher Education.* 22, 1 (2021), 63–75.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.* (2020).

[16] 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.

[17] 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), ep275.

[18] 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.

[19] 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.

[20] 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.

[21] Divate, M. and Salgaonkar, A. 2017. Automatic question generation approaches and evaluation techniques. *Current Science.* (2017), 1683–1691.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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).

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Krathwohl, D.R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice.* 41, 4 (2002), 212–218.

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

[36] 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 (2020), 121–204.

[37] 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 (2020), 121–204.

[38] 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, (2021), 101151.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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 (2021), 209–220.

[44] Moore, S., Nguyen, H.A. and Stamper, J. 2022. Leveraging Students to Generate Skill Tags that Inform Learning Analytics. Leveraging Students to Generate Skill Tags that Inform Learning Analytics (2022), 791–798.

[45] 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.

[46] Moore, S., Stamper, J., Bier, N. and Blink, M.J. 2020. A Human-Centered Approach to Data Driven Iterative Course Improvement. International Conference on Remote Engineering and Virtual Instrumentation (2020), 742–761.

[47] 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 (2022), 259–262.

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

[49] 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.

[50] 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, 225 (2018), 225.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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* (2022), 234–245.

[57] 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* (2021), 221–233.

[58] 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.

[59] 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.

[60] 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.

[61] 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).

[62] 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.