Examining the Effects of Student Participation and Performance on the Quality of Learnersourcing Multiple-Choice Questions

Steven Moore
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
StevenJamesMoore@gmail.com

Huy Anh Nguyen
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hn1@cs.cmu.edu

John Stamper
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
john@stamper.org

ABSTRACT

While generating multiple-choice questions has been shown to promote deep learning, students often fail to realize this benefit and do not willingly participate in this activity. Additionally, the quality of the student-generated questions may be influenced by both their level of engagement and familiarity with the learning materials. Towards better understanding how students can generate high quality questions, we designed and deployed a multiple-choice question generation activity in seven college-level online chemistry courses. From these courses, we collected data on student interactions and their contribution to the question-generation task. A total of 201 students enrolled in the courses and 57 of them elected to generate a multiple-choice question. Our results indicated that students were able to contribute quality questions, with 67% of them being evaluated by experts as acceptable for use. We further identified several student behaviors in the online courses that are correlated to their participation in the task and the quality of their contribution. Our findings can help teachers and students better understand the benefits of student-generated questions and effectively implement future learnersourcing activities.

Author Keywords

Question Generation; Learnersourcing; Student Participation; Online Education; Multiple-Choice Questions

CSS Concepts

• **Applied computing~Education~Interactive learning environments** • Applied computing~Physical sciences and engineering~Chemistry • Human-centered computing~Collaborative and social computing~Collaborative and social computing theory, concepts and paradigms~Collaborative content creation

L@S'21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06...$15.00.
https://doi.org/10.1145/3430895.3460140

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 [14,31]. 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 [7,18]. 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 [28,32,41]. In addition to evaluating student knowledge in both low-stakes and high-stakes environments, MCQs offer a scalable and equitable means of assessment [37]. The need for such scalability in assessment continues to increase, as class sizes continue to grow and more educational materials shift to being online [15]. 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 [24,35]. 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 [20,47]. Learnersourcing is a form of crowdsourcing in which students contribute novel content for future learners while engaging in a meaningful learning experience themselves [44]. While platforms like PeerWise [12], Quizzical [37], and RIPPLE [26] 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 [1]. 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 [41,48]. 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.

In this study, we specifically investigated the following three research questions:

RQ1: How does participation in online course materials impact students' contribution to learnersourcing activities?
RQ2: What is the quality of student-generated multiple choice questions from an optional and low-stakes learning environment?
RQ3: Are certain student interactions in an online course associated with the quality of a multiple-choice question they generate?

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.

BACKGROUND

Student Generation of Multiple-Choice Questions

The concept of learnersourcing is akin to crowdsourcing, but instead of enlisting the help of crowdworkers, students are leveraged to create or modify content while benefiting in a meaningful learning experience themselves [44]. This concept was formally proposed and defined in [27], where students were utilized to improve video content and interfaces during the learning process. Previous work by [47] has utilized learnersourcing techniques to have students generate MCQs and create knowledge links between questions. They 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 [33]. 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 [17,23]. One such popular system is PeerWise, which provides students with a custom learning environment for collaboratively generating and sharing questions [12]. 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 [16,32].

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 [11,13,16]. Another system, RiPPLE, also enables students to generate MCQs and formulate distractors for them, which requires them to think deeply about potential misconceptions [26]. 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. The present study involves students generating MCQs, but not through use of a separate system or in a way that requires their participation. Instead, students were presented with the activity directly embedded into an online course they were enrolled in, where they received no extra training and could choose to skip the question generation activity.

Assessing Multiple Choice Question Quality

Developing MCQs that cover the appropriate concepts and target higher cognitive levels can be challenging, even for expert instructors [15]. To assess the quality of MCQs, different item response theory and statistical methods have commonly been utilized [14,25]. 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 [8]. To help prevent these negative effects, previous studies have relied on qualitatively reviewing MCQs prior to testing them with students to confirm their validity [2,6,43]. 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 [19].

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 [6,38,43]. These studies often include an evaluation of the cognitive levels the MCQ assesses, which traditionally are recall and comprehension [41,49]. One particular study assessed the quality of over two thousand instructor-generated MCQs by utilizing a 19-criteria version of the IWF guidelines [42]. 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 [42]. However, while previous work focuses on applying the guidelines to instructor-generated questions, we apply them to student-generated ones.

Open Learning Initiative (OLI)

The Open Learning Initiative (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 [4]. It hosts instructional content and assessments, and is often being utilized as supplementary materials for blended learning like traditional online course platforms [21,40]. 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 types of assessments consist of a variety of types such as MCQs, short answer, and drop-down style questions. Detailed student interactions with the course materials, such as watching videos, answering questions, and accessing different pages are logged in the system. Specifically, 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 that question consists of three unique steps that each have their own set of feedback and correct or incorrect responses.

Previous research has demonstrated that the use of OLI in comparison to a traditional face-to-face classroom setting has led to a positive effect on the students' academic achievement [30]. They found that the use of an OLI statistics course led to improved learning outcomes in less than half of the time of their peers in a traditional learning environment. Another study by [5] demonstrated evidence of a positive effect on using OLI for blended learning and the persistence of students to remain enrolled in an introductory course. Students who interacted with the online materials were more likely to remain enrolled in the course and complete it than their peers who did not use the system. Ensuring students are both learning and participating is essential for systems like OLI, as overall engagement with activities in online courses is often low [45]. The present study utilizes an introductory chemistry course deployed on the OLI platform, where we elicited students to participate in an optional MCQ generation task.

METHODOLOGY

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 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 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 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.

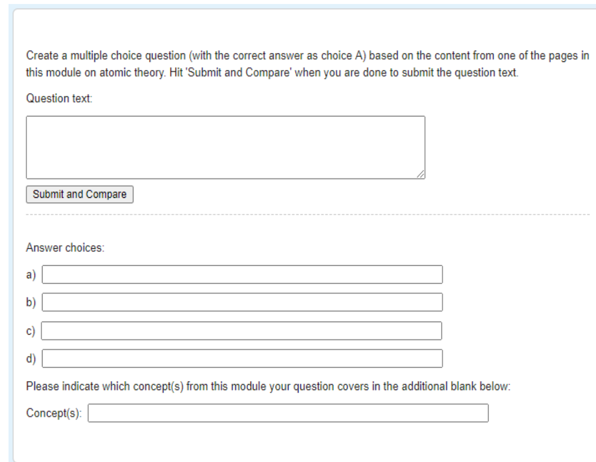

Figure 1. The MCQ generation question presented to students.

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 spent 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 [9]. 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 corrects and one incorrect for that step.

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 [19]. 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 [6,10,36,42,43]. A full description of the 19 items that make up the rubric can be found in Appendix A. 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 taxonomy that MCQs have been evaluated under in previous studies [28,34,38,39,42]. 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.

Table 2 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 2. 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.

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. Towards RQ1, 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 students interactions with the OLI materials and their participation with the MCQ generation task. For RQ2, we use measures of central tendency to report the varying IWFs and cognitive levels of the student-generated MCQs. We also include a Mann-Whitney U-test for determining if there is a significant difference for students that generated K2 questions instead of K1. Finally, to investigate RQ3 we use a series of unpaired t-tests to see which features of student behaviour 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 [3].

RESULTS

Student Participation

***RQ1**: How does participation with online course materials impact a student contribution to a learnersourcing activity?* 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 on 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 $M_1 = 32.29$ $(SD_1 = 27.17)$ minutes working on other activities in the course while students that did not do the task spent an average of $M_2 = 13.03$ $(SD_2 = 18.47)$ 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.

Question Quality

***RQ2**: What is the quality of a student-generated multiple-choice questions from an optional and low-stakes learning environment?* 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 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 [42]. 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 4. 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. 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 4. 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 5, 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 5. 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 6 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 7 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 7 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 6. 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 7. Two student generated MCQs evaluated by experts as being unacceptable due to their Item-writing Flaws.

Student Interaction and Question Quality

***RQ3**: Are students' behaviors in an online course associated with the quality of a multiple choice question they generate?*

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 ($M_{seconds} = 153$) and the quality of the contribution was found to not be statistically significant, $t(55) = 0.4769, p = 0.6353$).

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 [17]. While past MCQ generation methods have relied on external systems [12,26,37,46] or embedding the task in a high-stakes required assessment such as an exam [22], 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 [6,42,48]. Additionally, several student-generated MCQs did reach the K2 level of application and analysis and had significantly lower IWFs at this cognitieve 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 [29]. 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 [10,43]. 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 [26]. 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 low-stakes 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 [33]. 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 [29,44]. 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 [41,48]. Providing question writing guidelines for students could potentially help them construct MCQs at a more advanced cognitive level.

LIMITATIONS AND FUTURE RESEARCH

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.

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 behaviour that leads to a higher quality contribution for future learnersourcing tasks.

ACKNOWLEDGMENTS
The research reported here was supported in part by a training grant from the Institute of Education Sciences (R305B150008). Opinions expressed do not represent the views of the U.S. Department of Education.

REFERENCES
[1] Ester Aflalo. 2018. Students generating questions as a way of learning. *Active Learning in Higher Education*: 1469787418769120.

[2] Nadia Amini, Nicolas Michoux, Leticia Warnier, Emilie Malcourant, Emmanuel Coche, and Bruno Vande Berg. 2020. Inclusion of MCQs written by radiology residents in their annual evaluation: innovative method to enhance resident's empowerment? *Insights into Imaging* 11, 1: 1–8.

[3] Richard A. Armstrong. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* 34, 5: 502–508.

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

[5] William G. Bowen, Matthew M. Chingos, Kelly A. Lack, and Thomas I. Nygren. 2014. Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management* 33, 1: 94–111.

[6] Jared Breakall, Christopher Randles, and Roy Tasker. 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: 369–382.

[7] Andrew C. Butler. 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: 323–331.

[8] Sandra L. Clifton and Cheryl L. Schriner. 2010. Assessing the quality of multiple-choice test items. *Nurse Educator* 35, 1: 12–16.

[9] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4: 253–278.

[10] Tina Danh, Tamara Desiderio, Victoria Herrmann, Heather M. Lyons, Frankie Patrick, Gwendolyn A. Wantuch, and Kamila A. Dell. 2020. Evaluating the quality of multiple-choice questions in a NAPLEX preparation book. *Currents in Pharmacy Teaching and Learning*.

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

[12] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the Fourth international Workshop on Computing Education Research* (ICER '08), 51–58. https://doi.org/10.1145/1404520.1404526

[13] Paul Denny, Ewan Tempero, Dawn Garbett, and Andrew Petersen. 2017. Examining a student-generated question activity using random topic assignment. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, 146–151.

[14] David DiBattista and Laura Kurzawa. 2011. Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning* 2, 2: 4.

[15] Mercedes Douglas, Juliette Wilson, and Sean Ennis. 2012. Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International* 49, 2: 111–121.

[16] Denis Duret, Rob Christley, Paul Denny, and Avril Senior. 2018. Collaborative learning with PeerWise. *Research in Learning Technology* 26.

[17] Alireza Farasat, Alexander Nikolaev, Suzanne Miller, and Rahul Gopalsamy. 2017. Crowdlearning: Towards collaborative problem-posing at scale. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 221–224.

[18] Rebecca Grainger, Emma Osborne, Wei Dai, and Diane Kenwright. 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: 19–24.

[19] Thomas M. Haladyna, Steven M. Downing, and Michael C. Rodriguez. 2002. A review of multiple-choice item-

writing guidelines for classroom assessment. *Applied measurement in education* 15, 3: 309–333.

[20] Judy Hardy, Simon P. Bates, Morag M. Casey, Kyle W. Galloway, Ross K. Galloway, Alison E. Kay, Peter Kirsop, and Heather A. McQueen. 2014. Student-Generated Content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education* 36, 13: 2180–2194. https://doi.org/10.1080/09500693.2014.916831

[21] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4: 470–497.

[22] Alice Huang, Dale Hancock, Matthew Clemson, Giselle Yeo, Dylan J. Harney, Paul Denny, and Gareth Denyer. 2020. Selecting Student-Authored Questions for Summative Assessments. *bioRxiv*.

[23] Hyoungwook Jin, Minsuk Chang, and Juho Kim. 2019. SolveDeep: A System for Supporting Subgoal Learning in Online Math Problem Solving. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

[24] Jennifer A. Jones. 2019. Scaffolding self-regulated learning through student-generated quizzes. *Active Learning in Higher Education* 20, 2: 115–126.

[25] Ahmad Zamri Khairani and Hasni Shamsuddin. 2016. Assessing Item Difficulty and Discrimination Indices of Teacher-Developed Multiple-Choice Tests. In *Assessment for Learning Within and Beyond the Classroom*. Springer, 417–426.

[26] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics* 6, 3: 91–105.

[27] Juho Kim. 2015. Learnersourcing: improving learning with collective learner activity. Massachusetts Institute of Technology.

[28] Myo-Kyoung Kim, Rajul A. Patel, James A. Uchizono, and Lynn Beck. 2012. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American journal of pharmaceutical education* 76, 6.

[29] Josh B. Kurtz, Michael A. Lourie, Elizabeth E. Holman, Karri L. Grob, and Seetha U. Monrad. 2019. Creating assessments as an active learning strategy: what are students' perceptions? A mixed methods study. *Medical education online* 24, 1: 1630239.

[30] Marsha Lovett, Oded Meyer, and Candace Thille. 2008. The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. *Journal of Interactive Media in Education*.

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

[32] Heather A. McQueen, Cathy Shields, D. J. Finnegan, J. Higham, and M. W. Simmen. 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: 371–381.

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

[34] Badr Muhammad Moeen-uz-Zafar Khan. 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: 39.

[35] Eugenia MW Ng. 2014. Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring. *Computers & education* 73: 141–148.

[36] Adam Pate and David J. Caldwell. 2014. Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning* 6, 1: 130–134.

[37] C. Daniel Riggs, Sohee Kang, and Olivia Rennie. 2020. Positive Impact of Multiple-Choice Question Authoring and Regular Quiz Participation on Student Learning. *CBE—Life Sciences Education* 19, 2: ar16.

[38] Bonnie R. Rush, David C. Rankin, and Brad J. White. 2016. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education* 16, 1: 1–10.

[39] Ajit K. Sachdeva. 1996. Use of effective questioning to enhance the cognitive abilities of students. *Journal of Cancer Education* 11, 1: 17–24.

[40] Sandra Sanchez-Gordon and Sergio Luján-Mora. 2016. How could MOOCs become accessible? The case of edX and the future of inclusive online learning.

[41] Darina Scully. 2017. Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation* 22, 1: 4.

[42] Marie Tarrant, Aimee Knierim, Sasha K. Hayes, and James Ware. 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today* 26, 8: 662–671.

[43] Marie Tarrant and James Ware. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education* 42, 2: 198–206.

[44] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 1–10.

[45] Abeer Watted and Miri Barak. 2018. Motivating factors of MOOC completers: Comparing between university-affiliated students and general participants. *The Internet and Higher Education* 37: 11–20.

[46] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, 379–388.

[47] Iman Yeckehzaare, Tirdad Barghi, and Paul Resnick. 2020. QMaps: Engaging Students in Voluntary Question Generation and Linking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[48] Nikki L. Bibler Zaidi, Karri L. Grob, Seetha M. Monrad, Joshua B. Kurtz, Andrew Tai, Asra Z. Ahmed, Larry D. Gruppen, and Sally A. Santen. 2018. Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work? *Academic Medicine* 93, 6: 856–859.

[49] Alex Y. Zheng, Janessa K. Lawhorn, Thomas Lumley, and Scott Freeman. 2008. Application of Bloom's Taxonomy Debunks the"MCAT Myth." *SCIENCE-NEW YORK THEN WASHINGTON*- 319, 5862: 414.

Appendix A. The rubric of 19 item-writing flaws used to evaluate the student-generated multiple-choice questions.

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