Put the Students to Work: Generating Questions with Constructive Feedback

Richard Glassey
KTH Royal Institute of Technology
Stockholm, Sweden
glassey@kth.se

Olof Bälter
KTH Royal Institute of Technology
Stockholm, Sweden
ob1@kth.se

***Abstract*—Feedback has been long understood to be a vital component of learning. This is even more important in an online environment, where access to teachers is limited or unavailable. When online learning material is interspersed with opportunities for students to test their knowledge, there is a golden opportunity to give feedback and fix misconceptions immediately. All too often this opportunity is missed or mishandled in online learning platforms - either providing no feedback, or providing it at a coarse level of granularity. Furthermore, generating constructive feedback at a fine level of granularity is so time consuming as to be prohibitive. In this work, we build upon literature that has established students can be recruited to generate quality multiple choice questions, by exploring if students can also generate quality feedback to compliment the questions they generate. In the first iteration of an introductory programming course, we tasked 35 students to generate six questions each over three weeks, and analysed this dataset to generate principles of creating good multiple choice questions with constructive feedback. In the second iteration of the course, we repeated the same task, but provided students with the principles. The results showed an increased occurrence of good feedback when compared to the previous iteration with minimal need for additional intervention by teachers.**

***Index Terms*—Question-based learning, Online teaching, Feedback**

I. Introduction

It has been shown that online support for learning based on multiple choice questions (MCQ) with feedback can be very effective. Lovett, Meyer, and Thille showed in a randomised case-control study that learning time could be reduced by 50%, with similar examination results achieved when compared to traditional course offerings [1]. This online learning environment depends heavily upon questions with constructive feedback, interspersed throughout the learning material. Engaging with such environments, students immediately receive specific information on their (mis)understanding of the material, without the need for teacher intervention. However, generating a sufficient amount of MCQs addressing misconceptions with plausible answering alternatives and constructive feedback is very time consuming. In particular, the development of specific feedback for both correct and incorrect answering alternatives is often overlooked (or not supported by the learning environment itself) despite the contribution towards helping a student understand why they got a question wrong, without deliberately pointing to the correct answer.

To tackle this challenge, we explored PeerWise [2]. PeerWise has been successfully used to make students generate questions and answering alternatives in several previous studies, in different teaching contexts, e.g in Physics [3] and Biology [4], but we extend this here by also demanding that students generate constructive answering-alternative-dependent feedback as well as part of their task. The user interface supports an accompanying explanation field that can be used by students to add more information to students answering their question, such as more explanation, discussion about the misconception, or pointers to further study resources on the topic of the question.

We used two iterations of an introductory course in programming with 35 students per iteration, generating 440 questions in total. Initially, students were quite good at creating questions, somewhat less at providing reasonable answering alternatives, but much worse at providing constructive feedback. We then analyzed the question-answering alternatives-feedback triad provided by the students independently and when disagreement occurred we used the following discussion as a basis for formulating principles for good questions, answering alternatives and feedback. By providing these principles to the students, the quality of feedback improved throughout the remainder of the course. A further iteration of the course used the principles from the beginning and corrected both the cold-start problem of students giving feedback, and also an overall increase in feedback rated as good, and a reduction in feedback as rated as poor by academic evaluators. A final finding was that the presence of good feedback almost always predicts a good MCQ, making it an efficient filter for identifying high quality questions quickly.

As more engineering and computing courses use online learning environments, there is an opportunity to make course materials more effective to support student learning. The process and principles in this work contribute towards this effort by offering course developers an efficient and economical approach that does not compromise the effectiveness of the learning materials produced.

II. Background

Effective online learning that is both efficient (in terms of time cost for students) and economical (in terms of development/delivery cost for teachers) is a desirable goal. In this

section, we focus on how effective and efficient learning has been shown to be achievable in the Open Learning Initiative (OLI). However, there still exists a development cost for creating high quality questions with appropriate feedback that helps students correct their misconceptions. One possibility is to use student-generated content to reduce this cost. PeerWise is one such system that enables students to create multiple choice questions, but whilst it has been shown that students can create good questions, less is known about their ability to produce constructive feedback, a critical success factor for the OLI platform.

A. Question-based learning in OLI

In 2008, the Open Learning Initiative (OLI) at Carnegie Mellon University showed in a randomized case-control study it was possible to reduce the learning time in a university course in statistics with 50%, with maintained (or improved) results on midterms and finals [1]. This online environment depends heavily on questions with constructive feedback interspersed throughout the learning material. The questions improved learning by activating students. However, generating quality questions is time consuming. The Raccoon Gang estimated that it takes 90240 hrs to produce one hour of ready online learning content [5]. With limited resources available it is natural to turn to unpaid student work to generate the questions, but there are also pedagogical reasons to do this, as question writing can be an effective study and learning technique, leading to higher performance in assessments from those who author questions over those who do not [6].

Besides the benefits of online question-based learning, there are several studies indicating the advantages of the methodology: learning more material and achieving higher completion rates (99% vs. 41%) [7], high learning gains [8], learning benefits from the activities in OLI sixfold of only reading and watching videos [9] (who in their last study of four different courses had data from 12,500 students [10]).

A central part in OLI is the formative questions interspersed with the learning material. Answering these questions results in a reinforcement of either why the selected answer is correct, or if it was incorrect, how to think in order to understand what the correct answer would be. Ideally, these questions should be posed around misconceptions that students (usually) have and through wrong answers the students become aware of their misconceptions. The feedback guides them towards a correct(er) understanding of the world. Normally, this feedback is based on multiple choice questions, which limits the computer-generated feedback to areas where there are right and wrong answers. This might seem limiting, but the methodology has been used for courses in among other areas arts and humanities, social science and languages (see https://oli.cmu.edu/courses/). For a more elaborate explanation on OLI, see [11]; the underlying learning model [12]; analytics included in the learning dashboard [13]; and the use of learning curves to understand student mastery of skills and learning objectives [14].

Thus, it becomes essential to formulate good formative questions, ideally targeting students misconceptions, with constructive feedback that leads the students forward, but this is not an easy task and therefore very time consuming. Hence our effort to put the students to work on this, both for pedagogical reasons, but also labor saving for the teachers.

B. Generating MCQs using PeerWise

PeerWise is an online tool to enable students to create and answer MCQs [2]. It has been used by over 1500 academic institutions and has been the subject of many research articles. It belongs to the wider family of student-generated content, where responsibility for creating learning material of different varieties is delegated to the students themselves [15].

At its core, it is a simple web-based system that allows students to register, join different courses, author questions, answer questions generated by other students, as well as leave ratings of quality and difficulty and feedback on questions. Peerwise also employs gamification strategies, and students can accumulate points and badges for different activities, such as answering, giving feedback, sustaining runs of activity and so on. For questions that receive feedback from students who have answered a question, the system allows students to improve the original question, which creates the opportunity for iterative improvement of the set of questions over time. Research indicates that when PeerWise has been included in a course, students typically answer above and beyond the number of questions that is expected of them [16].

A common concern is whether students can create questions that have sufficient quality to be useful to be used as learning material. Denny, Luxton-Reilly & Simon conducted an analysis of student generated questions for an introductory programming course, where over 600 questions had been created [17]. They found a positive correlation (0.54) between their own ratings and the ratings that students had submitted. Furthermore, in terms of quality of the questions sampled, 93% were judged as having good clarity, 80% being free of errors, and 87% having feasible distractors/answering alternatives. The PeerWise interface enables students to provide an explanation when creating a question to further help students that have answered a question understand the topic. However, in terms of explanation/feedback, the authors only judged 25% as having explained the correct answer as well as providing helpful information to resolve the misconception.

Similarly, Bates, Galloway & McBride found that students in an introductory physics course could generate quality questions [3]. They categorised questions according to Blooms Taxonomy, and found that the majority of questions were identified as *Apply* or *Analyse*, whereas few questions were identified in the lower level (*Remember* and *Understand*) or higher level (*Evaluate* and *Create*). They also investigated the quality of explanations finding that about 45% of questions achieved a rating of good, and around 15% achieved a rating of excellent. However this was largely attributed to the substantial scaffolding approach used to train students in how to write questions: (1) orientation in MCQ terminology; (2) advantages

and disadvantages, (3) a self-diagnosis quiz, (4) a template encouraging students to aim for questions just outside their current level of understanding, and (5) exemplar MCQs with very high bars of creativity and complexity).

C. Can Students generate MCQs with good feedback?

Creating a MCQ is still mostly a creative act; as Ebel commented: "Each item as it is being written presents new problems and new opportunities. Just as there can be no set formulas for producing a good story or a good painting, so there can be no set of rules that will guarantee the production of good test items" [18]. Haladyna & Downing endeavoured to create a taxonomy of 43 multiple choice-item (henceforth MCQ) writing rules after conducting an analysis of 46 authoritative sources of educational measurement literature, from the time period of 1935 to 1989, that might assist in their creation [19]. This was motivated by the observation that whilst MCQs have been long considered an important component of educational measurement, the body of knowledge surrounding their construction was not well established nor formalised in any framework or theory. In Haladyna, Downing & Rodriguez [20] the list of writing rules was further refined to 31 after reviewing what was found in 27 textbooks on educational testing and the results of 27 research studies and reviews published since 1990. The top most items of consensus included: (1) Central idea in stem/question, (2) Avoid clues, (3) Make distractors/answering alternatives plausible, (4) Use novel material, and (5) Keep the length of choices/answering alternatives about equal.

However, none of the rules concern the provision of feedback that might be given to the student upon answering a question, only the question and the answering alternatives. As the previous section has shown, students can create good quality MCQs, as judged by academic evaluators. There is some evidence that students can also provide good or excellent explanations, but not to the same level or consistency as for the questions and answering alternatives. The central question of this work is whether students can construct good MCQs that also contain constructive feedback to help address the misconceptions that are being targeted. The next section will describe the context for our work, how students were motivated to create MCQs as part of their assessment, and how the generated questions were analysed to help identify guiding principles for students when creating MCQs with constructive feedback.

III. Method

The Software Development Academy (SDA) is an intense three month training for mature students that runs twice a year at the KTH Royal Institute of Technology (Stockholm, Sweden. It consists of a series of modules that build the necessary skills for a junior Java developer job-hunting within the local IT industry. The training begins with a three week module on introductory programming. In the sixth iteration (SDA-6), it was decided to remove the end of module exam as the main form of assessment, and replace it with a question generation task using the PeerWise system (see Sec. II for more details on its features).

A. Part One: Generating Questions by Students

At the end of each teaching week, students ($n = 35$) were instructed to create two MCQs, based on different topics that had been covered within the week. In the first session, students were introduced to the PeerWise system, which was populated with ten sample questions taken from the module content they had already encountered in the OLI platform. Once students had successfully logged into the system, they were instructed to create their questions. They were free to answer, comment and rate as many questions as they liked. The minimal requirement to pass the module was to generate six questions by the end of the final week. Students were given the following advice before starting their task:

1) The question topic and difficulty is inline with the module content.
2) The question is clear and unambiguous.
3) There are no accidental typos.
4) The possible answers are reasonable alternatives so there is a challenge.
5) The feedback describes why each alternative was incorrect, but does not reveal the answer in doing so.

In terms of further intervention, an academic reviewed the first weeks questions and posted individual feedback where the students had deviated from the original instructions. During the second weeks question generation session, an oral summary of the general themes emerging from the review of questions was given prior to the second attempt in order to improve the overall question quality. This was also repeated before the final session. Otherwise students were free to interact with the PeerWise system as much or as little as they wanted throughout the module.

B. Part Two: Analysing the Questions

Analysis of the questions was divided into two phases. In the initial analysis phase, a random sample of ten questions was taken each week, and were assessed by two academics working together on the task. This served the purpose of devising and calibrating a common method for judging MCQs, and to resolve any differences of opinion regarding the quality of the question, answering alternatives and feedback. It should be noted that a single MCQ is composed of these three components. A simple scoring system was devised for appraising each component of a question, shown in Table I:

In the exhaustive analysis phase, we randomly divided two thirds of the questions between ourselves (creating an overlap for assessing levels of inter-academic agreement), and graded each MCQ component based on the scoring system devised in the initial analysis. This selection of questions was based upon the observation that it is sometimes difficult to judge whether it is the question, or the answering alternatives that contribute most to poor quality, i.e. the quality of answering alternatives is dependent on the question, and the feedback is dependent upon both the question and the answering alternatives. When

TABLE I
RATING SCALE FOR QUESTION, ANSWERING ALTERNATIVE AND FEEDBACK COMPONENTS OF MCQS

2	Good	No changes are needed, besides formatting and language corrections.
1	Ok	Minor changes are needed to correct and elevate to a good rating.
0	Poor	Major changes are needed to fix the question or it is beyond repair.

the scores differed by more than one step, either for one part or the combination of all three parts, we had a discussion on why we disagreed, and what principle we could formulate and agree upon in order to judge similar questions in a more consistent and coherent way. The purpose of formulating principles is to convey quality aspects of MCQs to students in the subsequent iterations of the SDA programme in order to improve the quality of the questions they generate. The next section describes the results of this approach over two iterations of the SDA programme, as well as the formulation and impact of the principles of a good MCQ.

IV. RESULTS

A. Generation of Principles

The initial instructions given to the students contained several candidates for principles for good MCQs. The first one concerns the domain: (P1) all questions should be in the relevant domain (e.g. the question should address what is being taught that week). The second is based on guiding principles of the OLI approach: (P2) all answering alternatives should ideally address misconceptions that students have and should be plausible, e.g. avoid humour or impossible answers. Whilst it is challenging for a student to know which misconceptions other students have in general, they can identify their own misconceptions encountered during their own learning experience. However, despite this starting point we wanted to explore the dataset of questions generated by students to learn the patterns and problems that could be addressed in advance by having a comprehensive yet compact set of guiding principles at hand.

The first finding repeats what other researchers have found: students engage positively with the PeerWise system and the task of generating questions [16]. In total, 230 questions were generated over three weeks, twenty more than expected. Furthermore, whilst students were only instructed to author questions and then use the system as much or little as they wanted, 50% of the students answered 100 or more questions, and 75% of the students answered 50 or more questions, see Fig. 1. Finally, whilst anecdotal, students verbally reported how much they liked the system, typically preparing their questions in advance of the dedicated session each Friday, and even the least active student in Fig. 1 managed to generate the required six questions with minimal encouragement.

The second finding, when exploring a random sample of 10 questions per week, was that students did not struggle to generate good questions and answering alternatives, but they struggled to produce good feedback. However, with encouragement to provide better feedback, improvements were observed both in the second and third weeks, see Fig. 2. The following two examples give a sense of what is considered poor feedback and how it compares with good feedback. In the poor example, it is hard to comprehend the written text, and it only provides a single overall explanation. In contrast, the good example is both readable without being too verbose, and addresses each answering alternative with a unique explanation that does not reveal the correct answer:

Example of Poor feedback:

> *Here we have a sufix which first print or shows the value and then operator works. For prefix first the operation is done.*

Example of Good feedback:

> *A: Correct: Autoboxing is what happens when the compiler automatically wraps a primitive value in an appropriate wrapper object*
> *B: Close: Unboxing is the flipside of autoboxing that a compiler does automatically when a value is retrieved and stored as a primitive*
> *C: Type conversion is something that can be done within Java (and that might help us store primitive values with some effort), but this isn't quite right*
> *D: There's actually no particular syntax involved in autoboxing and unboxing: they're simply done by the compiler behind the scenes*

Despite being a small sample, several potential principles for good feedback emerged in discussions throughout the initial analysis. Students should produce (P3) unique feedback per answering alternative (already suggested to the students in their instructions), which is (P4) constructive in helping the student overcome their misconception, and (P5) does not reveal the correct answer. This cluster of principles was the most frequent reason for a poor score occurring for the feedback component. Another recurring problem was that the formulation of the question, answering alternatives or feedback impeded readability to the point where a basic question with bad formulation became much more difficult than it should have been, e.g. the answering alternatives were too long to easily hold in memory when trying to answer the question, or the use of negation was misleading when reading the question.

The third finding underlines that it was possible to improve the quality of feedback when the entire dataset was considered, from 34.2% of MCQs in week 1 scoring good, increasing to 43.7% MCQs scoring good in week 3, as shown in Table II. However, reduction in scores for question and answering alternatives reflect the challenge that students face writing MCQs as the course material becomes progressively more advanced. But as this was conducted over a short time scale, it is hard to rule out novelty effects in using PeerWise, size of sample (number of students and generated MCQs), and

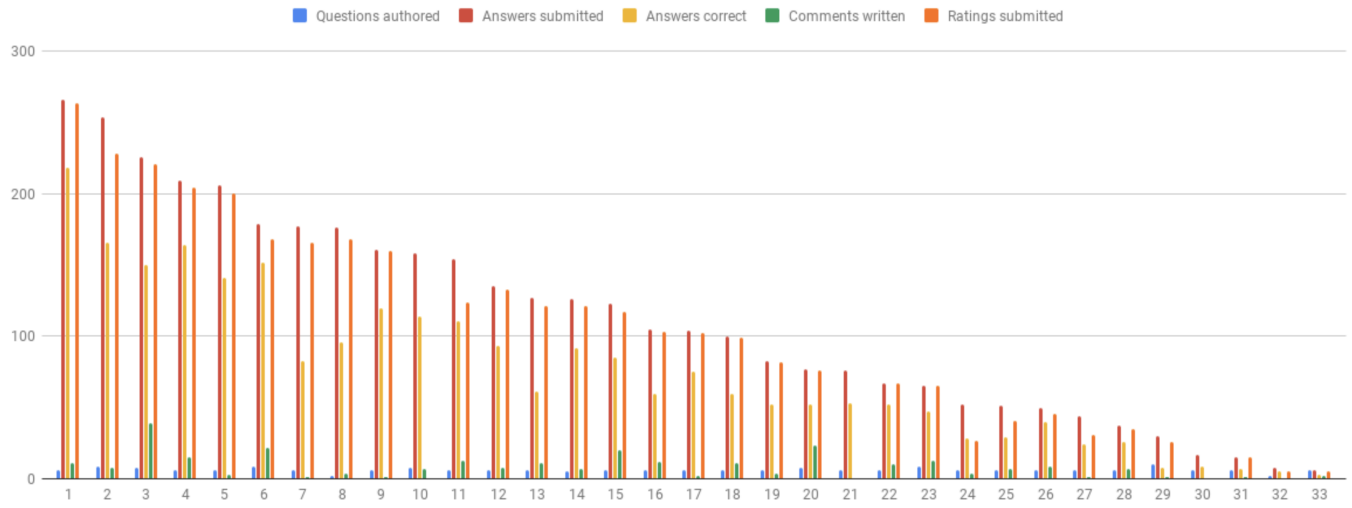

Fig. 1. Participation analysis with PeerWise, where each cluster of bars represents an individual student.

Identifier	Question	Alternatives	Feedback
4002755	2	1	0
4006548	2	2	0
4002821	1	0	0
4002772	2	2	0
4002802	2	2	0
4002774	2	2	0
4002783	1	1	0
4002837	2	2	2
4002763	0	0	0
4002749	2	1	2
Sum	**16**	**13**	**4**

Week 1

Identifier	Question	Alternatives	Feedback
4013096	2	2	0
4013074	2	2	2
4013130	2	2	2
4013068	2	2	2
4013213	0	1	1
4013120	2	2	2
4016137	2	0	1
4016160	0	1	0
4013255	2	2	2
4019041	2	2	1
Sum	**16**	**16**	**13**

Week 2

Identifier	Question	Alternatives	Feedback
4026156	2	1	0
4031157	2	0	0
4025792	2	1	2
4026903	2	2	2
4025824	2	2	2
4029801	2	2	2
4025845	2	1	1
4025814	0	1	0
4029659	2	2	2
4026159	2	2	2
Sum	**18**	**14**	**13**

Week 3

Fig. 2. Random sample of student generated questions across three weeks of SDA-6, showing the progression of scores in terms of question, answering alternative and feedback quality, where the scores are 2 / Good, 1 / OK, and 0 / Poor.

inconsistent application of the nascent principles by ourselves influencing the results.

The exhaustive analysis of overlapping questions resulted in a further five potential principles, as it was possible to see a lot more variability in MCQs when analysing the complete dataset. Questions that were ok or poor highlighted other principles that were not provided to the students in the initial instructions: (P6) focus on a single misconception, (P7) avoid reference look-up questions (for the type of information you would normally search for in everyday work situations) and (P8) avoid the need for any external systems (e.g. a compiler) to answer a question. In terms of answering alternatives, (P9) use three or more alternatives, to avoid binary choice, true / false type questions, and (P10) ensure that the formatting maximises readability (e.g. if alternatives become hard to comprehend, it is difficult to hold them all in memory when answering the question, which leads to arduous re-reading multiple times).

From the original instructions, the initial analysis, and the exhaustive analysis, a set of potential principles for creating good MCQs was identified, organised into three categories corresponding to the components of a MCQ, and summarised in Table III for convenience. The next section will describe the validation of these principles in the following iteration of the SDA programme.

B. Validation of the Principles

The following iteration of the programme, SDA-7, afforded the opportunity to validate the principles with a new cohort of students (n=32). The process was the same as before, except the students were given the principles to read before creating their first week of questions. Rather than give students individual feedback on questions that did not meet our expectations, students were simply reminded about following the principles. Once the students had completed generating their questions, we repeated the same initial analysis as before on a random sample of 10 questions per week.

In the previous iteration, SDA-6, the cumulative score for feedback in week 1 was 4 (see Fig. 2). As shown here, for

TABLE II
PERCENTAGE OF QUESTIONS, ALTERNATIVES AND FEEDBACK PER WEEK FOR ALL STUDENT GENERATED MCQS IN SDA-6 (N=230).

	Question			Answering Alternative			Feedback		
	Good	Ok	Poor	Good	Ok	Poor	Good	Ok	Poor
Week 1	84.2%	13.2%	2.6%	61.8%	34.2%	3.9%	34.2%	57.9%	7.9%
Week 2	78.5%	12.3%	9.2%	78.5%	20.0%	1.5%	43.1%	38.5%	18.5%
Week 3	76.7%	19.9%	3.4%	59.2%	38.0%	5.6%	43.7%	47.9%	11.3%

TABLE III
SUMMARY OF THE PRINCIPLES OF GOOD MULTIPLE CHOICE QUESTIONS DERIVED FROM THE INITIAL AND EXHAUSTIVE ANALYSIS.

Principles of Good Questions	(1) Question is from the course domain.
	(2) Question is targeted towards a misconception.
	(3) Question is not based on reference look up.
	(4) Question is reasonable to solve without external systems.
Principles of Good Answering Alternatives	(5) Three or more answer alternatives are provided.
	(6) Answer alternatives are plausible and linked to the misconception.
	(7) Answer alternatives are formulated to maximize readability.
Principles of Good Feedback	(8) Feedback is constructive.
	(9) Feedback is unique and provided for each answer alternative.
	(10) Feedback for answer alternatives does not reveal the answer.

Identifier	Question	Alternatives	Feedback
4298249	2	2	1
4290631	2	2	2
4295434	2	2	0
4290611	2	1	1
4293777	2	2	2
4295402	2	2	2
4290509	2	0	1
4290516	2	2	0
4298252	2	2	0
4295392	2	2	2
Sum	**20**	**17**	**11**

Week 1

Identifier	Question	Alternatives	Feedback
4316359	2	2	0
4315610	2	0	0
4316760	1	2	0
4312013	2	2	2
4312507	2	2	1
4317957	2	2	0
4316160	2	2	2
4317222	2	2	2
4314600	1	1	2
4311335	2	2	2
Sum	**18**	**17**	**11**

Week 2

Identifier	Question	Alternatives	Feedback
4327992	1	1	2
4327998	2	2	2
4321662	2	2	1
4325304	2	2	0
4328016	1	2	1
4327301	2	2	2
4322569	1	1	0
4328040	2	2	2
4322643	2	2	1
4327315	1	0	0
Sum	**16**	**16**	**11**

Week 3

Fig. 3. Random sample of student generated questions across three weeks of SDA-7, showing the progression of scores in terms of question, answering alternative and feedback quality, where the scores are 2 / Good, 1 / OK, and 0 / Poor.

SDA-7 the sum was raised to 11 and subsequently sustained for the following two weeks (see Fig. 3). The cold-start issue experienced in SDA-6, where students may not have thought there was much need to provide sufficient or good feedback, or any at all, appears resolved in this sample. Despite this improvement, there still appears to be a limit that cannot be broken through using the principles by themselves; the previous iteration maintained a sum of 13 in terms of feedback in the final two weeks, and in the current iteration shown in Fig. 3, a sum of 11 is maintained across the final two weeks.

The final result emerged from performing a comprehensive analysis of all questions generated in SDA-7 in terms of their feedback. We decided that there was no strong need to analyse the question nor answering alternatives. The reasoning was simple: hardly any questions generated by students with feedback rated as 2 / Good, had poor question construction or choice of answering alternatives. That is, presence of good feedback is an excellent proxy for finding good questions. A further motivation was practical - this decreased the time required to judge a large number of questions from a few hours

down to a single hour or less in our own anecdotal experience.

TABLE IV
COMPARISON OF QUALITY OF CONSTRUCTIVE FEEDBACK THAT STUDENTS PROVIDED WITH THEIR QUESTIONS BETWEEN SDA-6 (BOTH THE INITIAL ANALYSIS WITHOUT THE PRINCIPLES AND A RETROSPECTIVE ANALYSIS WITH THE PRINCIPLES) AND SDA-7 (WITH THE PRINCIPLES).

Iteration / Feedback	Good	Ok	Poor
SDA-6 Pre-principles	42.9%	45.8%	11.3%
SDA-6 Post-principles	32.4%	36.6%	31.0%
SDA-7 Post-principles	49.3%	33.8%	16.9%

Overall, there was a clear improvement in the percentage of usable questions with constructive feedback. In SDA-6 (pre-principles) shown in Table IV, we can see a lot of optimism in the questions judged as being 2 / Good or 1 / Ok, however, once a stricter interpretation of the principles was applied, SDA-6 (post-principles) shows a consistent shift downwards in usability, with the almost tripling of the percentage of poor questions. Looking at the distribution for SDA-7 (post-principles) there are positive results with almost half of the questions being considered good, with a further third being considered ok. Finally, using the principles in this evaluative manner creates a secondary use beyond helping to guide students to create better questions in the first place, which could help considerably in efforts to filter and discover the best questions for inclusion in future online learning material.

V. DISCUSSION

Returning to the central question of this work, can students create good MCQs with constructive feedback, we can answer yes, but with some limitations. First, given the opportunity to provide feedback for a MCQ, students will not perform very well with little or no guidance. Our own results agreed with one of the earliest studies of quality of MCQs that also considered the feedback component [17]. They expressed their disappointment that whilst students could create good MCQs, they overwhelmingly failed to provide any useful feedback that might help another student correct their misconceptions. However, in the first iteration, after one week, we found that by drawing attention to the importance of the feedback component, the results at least started to improve over the next two weeks. So there was a cold start situation which could be attributed to the students general lack of experience of providing this type of feedback, but it was amenable to positive change and improvement through teacher intervention.

Second, some training or guidance is required to generate improvements in feedback quality. In the second iteration of the course, we observed a marked improvement in providing constructive feedback when students were asked to read and consider applying the principles of good MCQs. This would seem to provide support for the idea that students can be motivated to provide constructive feedback. The extensive scaffolding used in Bates, Galloway & McBride [3] also led to better results in this regard. However, in terms of more economical approaches, we would argue that the guiding principles are a much lower cost solution that might be much more convenient than the scaffolding approach (see Sec. II).

A third limitation was the ceiling that we encountered for feedback quality. Whilst we only had three weeks of students using PeerWise, we found that improvements from the first to second week did not continue to increase in the third week. Two explanations emerge for this limit within the context of our work. Firstly, this can be in part caused by the lack of steering in PeerWise, which only had one textbox to use for all the feedback. Ideally we wanted students to create explanations for all answering alternatives for a question, but this might have felt artificial using a single textbox. This incongruence was mentioned early to the students and they were encouraged to suspend reality in that regard, but it undoubtedly would have had some effect - "why not just give one canonical explanation?". Secondly, when open discussions with students concerned the motivations to provide feedback, some felt the questions they had created did not require long explanations - more difficult questions were easier to write about, but simple concepts should just point to a reference, such as the textbook or other online learning material.

Stepping back, we took a bottom-up approach towards generating the principles from an analysis of the questions that students generated, rather than start with a top-down approach from advice contained in the literature. Previous works have extensively reviewed decades of ideas and advice on writing MCQs [19], [20] and devised rules (31 to be precise) that an academic should follow. In retrospect, perhaps we should have given the students these rules, however none of them consider feedback, and asking a student to adhere to 31 rules might be somewhat optimistic in terms of consistency and compliance across the cohort.

Instead, the identified principles are intended as instructions for students generating questions, which take into consideration the patterns and problems that students tend to adopt. Whether these principles are useful for teachers generating questions is a different matter. The first principle in each sections could be regarded as common sense, at least for teachers who are experienced in creating questions. However, these three principles were included in the instructions to the students, and still they (often) failed to follow them. It could be that the principles could be used as a checklist by teachers.

Somewhat amusingly, when we had the chance to ask teachers to apply our principles at an informal writing retreat, all teachers agreed with our principles as things they would naturally apply themselves, yet rather predictably from our experience with the students, the teachers also managed to create good MCQs, but completely failed to provide us with constructive feedback. As we discovered in our own exhaustive analysis of two sets of student generated questions, presence of quality feedback clearly identifies a good question. Perhaps both students and teachers could benefit from our principles, but ultimately someone has to check that they actually have been applied, and this remains a topic for future work.

In terms of limitations, we can ask if these principles are generalizable outside the area of computer science. For example the principles of not basing questions on reference lookup or use of external systems may be limiting for some academic areas that would like to base their questions on e.g. definitions or use of domain specific tools. Of course, if the skill of looking up a certain definition is important, it could still be a good question, and these are principles, not commandments. However, we would argue that even questions on definitions would improve if it is formulated so that you not only need to find the definition, but actually understand the implications of it.

Finally, we had a limited number of students and generated questions, and this is something we will endeavour to improve by continuing to repeat our approach with successive versions of the SDA programme, and to explore the idea of creating a better user interface that does not make it easy for students to skip the step of creating constructive feedback for each answering alternative.

VI. Conclusion

Effective and efficient online learning is greatly helped by having regular opportunities to test knowledge and receive constructive feedback in order to address misconceptions. Creating this learning material comes at a cost, however it is a cost that can be allieviated by recruiting students to assist with the task. We have performed a study that attempted to identify principles for how to generate good questions, answering alternatives, and constructive feedback. These principles were generated from the analysis of student generated questions, allowing us to target the advice towards the typical patterns and problems that emerged from that dataset. From this, we could then observe the improvements in the generation of constructive feedback when students were advised to apply the principles. Whilst the creation of efficient, effective and economical learning material may appear as a trillemma - *you can only pick two* - we suggest that there is a virtuous cycle to be found. Students can generate content, academics can provide principles or training aimed at students as authors, and students can benefit by engaging with better learning material. Not all learning materials produced will be perfect, but with careful guidance, that percentage can increase with relative ease.

References

[1] M. Lovett, O. Meyer, and C. Thille, "The open learning initiative: Measuring the effectiveness of the oli statistics course in accelerating student learning." *Journal of Interactive Media in Education*, 2008.

[2] P. Denny, A. Luxton-Reilly, and J. Hamer, "The peerwise system of student contributed assessment questions," in *Proceedings of the tenth conference on Australasian computing education-Volume 78*. Citeseer, 2008, pp. 69–74.

[3] S. P. Bates, R. K. Galloway, and K. L. McBride, "Student-generated content: Using peerwise to enhance engagement and outcomes in introductory physics courses," in *AIP Conference Proceedings*, vol. 1413, no. 1. American Institute of Physics, 2012, pp. 123–126.

[4] H. A. McQueen, C. Shields, D. Finnegan, J. Higham, and M. Simmen, "Peerwise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention," *Biochemistry and Molecular Biology Education*, vol. 42, no. 5, pp. 371–381, 2014.

[5] R. Gang, "How much does it cost to develop an online course?" https://raccoongang.com/blog/how-much-does-it-cost-create-online-course/, 2019, last visited 2020-02-12.

[6] P. W. Foos, "Effects of student-written questions on student test performance," *Teaching of Psychology*, vol. 16, no. 2, pp. 77–78, 1989.

[7] C. D. Schunn and M. Patchan, "An evaluation of accelerated learning in the cmu open learning initiative course logic & proofs," *Report, Learning Research and Development Center, University of Pittsburgh*, 2009.

[8] P. S. Steif and A. Dollár, "Study of usage patterns and learning gains in a web-based interactive static course," *Journal of Engineering Education*, vol. 98, no. 4, pp. 321–333, 2009.

[9] K. R. Koedinger, J. Kim, J. Z. Jia, E. A. McLaughlin, and N. L. Bier, "Learning is not a spectator sport: Doing is better than watching for learning from a mooc," in *Proceedings of the second (2015) ACM conference on learning@ scale*, 2015, pp. 111–120.

[10] K. R. Koedinger, E. A. McLaughlin, J. Z. Jia, and N. L. Bier, "Is the doer effect a causal relationship? how can we tell and why it's important," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 2016, pp. 388–397.

[11] C. Thille and J. Smith, "Cold rolled steel and knowledge: What can higher education learn about productivity?" *Change: The Magazine of Higher Learning*, vol. 43, no. 2, pp. 21–27, 2011.

[12] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive science*, vol. 36, no. 5, pp. 757–798, 2012.

[13] M. Lovett, "Cognitively informed analytics to improve teaching and learning," *Presentation at EDUCAUSE Sprint. Retrieved October*, vol. 5, p. 2015, 2012.

[14] O. Bälter, D. Zimmaro, and C. Thille, "Estimating the minimum number of opportunities needed for all students to achieve predicted mastery," *Smart Learning Environments*, vol. 5, no. 1, p. 15, 2018.

[15] S. Wheeler, P. Yeomans, and D. Wheeler, "The good, the bad and the wiki: Evaluating student-generated content for collaborative learning," *British Journal of Educational Technology*, vol. 39, no. 6, pp. 987–995, 2008.

[16] P. Denny, A. Luxton-Reilly, and J. Hamer, "Student use of the peerwise system," in *Proceedings of the 13th annual conference on Innovation and technology in computer science education*, 2008, pp. 73–77.

[17] P. Denny, A. Luxton-Reilly, and B. Simon, "Quality of student contributed questions using peerwise," in *Proceedings of the Eleventh Australasian Conference on Computing Education-Volume 95*, 2009, pp. 55–63.

[18] R. L. Ebel, "Writing the test item," *Educational measurement*, pp. 185–249, 1951.

[19] T. M. Haladyna and S. M. Downing, "A taxonomy of multiple-choice item-writing rules," *Applied measurement in education*, vol. 2, no. 1, pp. 37–50, 1989.

[20] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied measurement in education*, vol. 15, no. 3, pp. 309–333, 2002.