Research in Learning Technology
Vol. 29, 2021

ORIGINAL RESEARCH ARTICLE

Selecting student-authored questions for summative assessments

Alice Huang[a], Dale Hancock[a], Matthew Clemson[a], Giselle Yeo[a], Dylan Harney[a], Paul Denny[b], and Gareth Denyer[a]*

[a]*School of Life and Environmental Science, University of Sydney, Sydney, Australia;*
[b]*School of Computer Science, University of Auckland, Auckland, New Zealand*

(Received: 28 July 2020; Revised: 23 October 2020; Accepted: 11 December 2020; Published: 03 February 2021)

Production of high-quality multiple-choice questions (MCQs) for both formative and summative assessments is a time-consuming task requiring great skill, creativity and insight. The transition to online examinations, with the concomitant exposure of previously tried-and-tested MCQs, exacerbates the challenges of question production and highlights the need for innovative solutions. Several groups have shown that it is practical to leverage the student cohort to produce a very large number of syllabus-aligned MCQs for study banks. Although student-generated questions are well suited for formative feedback and practice activities, they are generally not thought to be suitable for high-stakes assessments. In this study, we aimed to demonstrate that training can be provided to students in a scalable fashion to generate questions of similar quality to those produced by experts and that identification of suitable questions can be achieved with minimal academic review and editing. Second-year biochemistry and molecular biology students were assigned a series of activities designed to coach them in the art of writing and critiquing MCQs. This training resulted in the production of over 1000 MCQs that were then gauged for potential by either expert academic judgement or via a data-driven approach in which the questions were trialled objectively in a low-stakes test. Questions selected by either method were then deployed in a high-stakes in-semester assessment alongside questions from two academically authored sources: textbook-derived MCQs and past paper questions. A total of 120 MCQs from these four sources were deployed in assessments attempted by over 600 students. Each question was subjected to rigorous performance analysis, including the calculation of standard metrics from classical test theory and more sophisticated item response theory (IRT) measures. The results showed that MCQs authored by students, and selected at low cost, performed as well as questions authored by academics, illustrating the potential of this strategy for the efficient creation of large numbers of high-quality MCQs for summative assessment.

Keywords: multiple choice questions; student authored questions; question banks; examinations; item response theory; peerwise

Introduction

The production of high-quality multiple-choice questions (MCQs) has always been a significant challenge for teachers. There is a constant demand for MCQs for both summative assessments and as resources for students to revise and engage with course

*Corresponding author. Email: gareth.denyer@sydney.edu.au

Citation: Research in Learning Technology 2021, **29**: 2517 - http://dx.doi.org/10.25304/rlt.v29.2517

content. The construction of appropriately aligned and pitched MCQs is not a trivial task. MCQs are often criticised for not assessing conceptual understanding and, instead, being overly focused on recall of subject-level minutiae (Biggs & Tang 2011). Indeed, a large-scale review of MCQs used across the United States in university biology courses, including molecular biology, revealed that more than 90% targeted the lowest two levels of the revised Bloom's taxonomy (Momsen *et al.* 2010). However, like all forms of assessment, it is possible to write MCQs that test deep understanding, extrapolation and other educational outcomes high on Bloom's Taxonomy (Harper 2003). To do this, the teacher must not only possess mastery of their subject material but also be aware of their students' insights and misconceptions.

The recent increased deployment of online examinations exacerbates the problem of MCQ supply and demand. Any material used in online assessments must be assumed to be in the public domain and, therefore, cannot be reused in subsequent assessments without risks to academic integrity. The danger of reusing questions on high-stakes exams has long been acknowledged (McCoubrie 2004), with recent empirical evidence revealing considerable deterioration in the psychometric properties of questions when reused over several years (Panczyk *et al.* 2018). This has particular impact on 'keeper' questions, those in which every experienced academic curates and re-uses in exams over several years, each time reflecting on the performance metrics of the item and perhaps subtly modifying to give ever-improved discrimination and power. The problem is further compounded by the fact that even the most thoughtfully constructed questions need to be validated in real assessment situations to confirm student interpretation, identify ambiguities and validate assumptions about difficulty and discrimination. In response to the issues above, it is not surprising that many academics are attracted to use pre-prepared and presumably field-tested question banks provided by textbook publishers and in online repositories. Yet, this is not a panacea, as even commercial question banks frequently contain items with flaws (Masters *et al.* 2001) and the questions themselves may not align well to bespoke syllabuses.

Some academics have explored the production of large pools of MCQs by their students through the use of crowdsourcing (Aflalo 2018; Amini *et al.* 2020; McLeod & Snell 1996). The resulting banks of student-generated questions are typically used for formative feedback and practice and often prove popular resources for study and exam revision (Duret *et al.* 2018; Gooi & Sommerfeld 2015; Papinczak *et al.* 2012; Walsh *et al.* 2018). A widely used tool for supporting such activities is PeerWise, an online platform where students can author and answer MCQs, as well as provide feedback on questions created by their peers (Denny *et al.* 2008; Denny, Luxton-Reilly & Hamer 2008).

Despite growing evidence supporting the pedagogical value of getting students to create questions for each other, most instructors would be reluctant to use the student-generated questions in formal, summative assessments. The quality of the questions that students produce can vary widely (Bottomley & Denny 2011; Purchase *et al.* 2010; Snow *et al.* 2019), and in addition to minor deficiencies in the clarity of wording and quality of plausible distractors, both Bates *et al.* in Physics (Bates *et al.* 2014) and Galloway and Burns in Chemistry (Galloway & Burns 2015) found that about 5% of student-authored questions were fundamentally incorrect.

Our own observation over several years of using PeerWise to promote engagement and reflection of course learning outcomes is that students are well placed to recognise dissonance between their own and their peers' insights, which can reveal misconceptions by either party. We were also aware of studies in which instructors proactively improved the quality of the questions produced by students by providing the class with MCQ-writing manuals covering structural and content elements (Jobs *et al.* 2013) and having students attend dedicated MCQ-writing tutorials before authoring their own questions (Bates *et al.* 2014). We therefore hypothesised that, with adequate coaching in the art of creating MCQs, and with allocation of learning outcomes to student-authors at a suitable level of granularity, we could leverage the class to produce a large bank of assessment-grade MCQs. We further hypothesised that it would be possible to screen for the most suitable student-designed questions using performance metrics from a broadly implemented low-stakes assessment.

Accordingly, we scaffolded activities over the semester to develop skills in authoring and critiquing MCQs, specifically training students to incorporate peer-confessed insights and misconceptions into question stems and distractors. Student-authored questions were then evaluated in two different ways to assess suitability for inclusion in a high-stakes assessment: (1) performing an objective, data-driven approach, by setting a low-stakes assessment and using performance data to identify questions with potential; and (2) taking a hypothesis-driven approach, by identifying candidate questions through academic review and editing. Student-authored questions from both sources were then pooled with textbook-derived MCQs and past paper questions, which have been used in previous years' exams. The performance of all these questions was then evaluated using traditional evidence-based metrics (difficulty and discrimination index), as well as more detailed techniques (item response theory and distractor analysis).

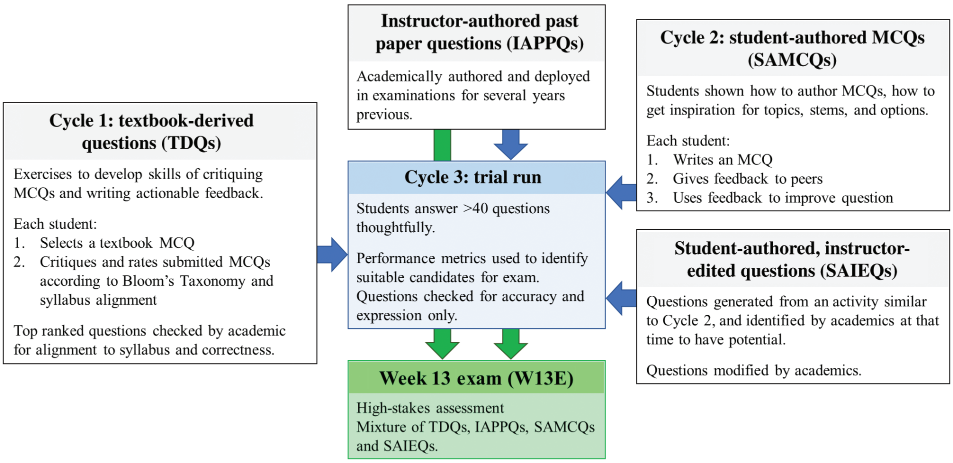
Methods

Ethics

Processes were conducted in accordance with the Sydney University Human Ethics protocol 'Investigating how engagement with peer-generated assessment impacts student success. Project number: 2017/131'. Under this protocol, students were able to anonymously, and without prejudice, object to their contributions being included through an online form set up by the University Research Office. An independent administrator, not involved with the study, performed the cross-checking.

Overview of activities

Activities for a second-year biochemistry cohort (over 600 students), were organised into three cycles (Figure 1), each with the outcome of generating or selecting questions for a summative, high-stakes, in-semester exam – the Week 13 examination (W13E). In Cycle 1, students were trained to dissect the intent and structure of existing MCQs, gain the skills to recognise attributes of strong and weak MCQs, and be empowered with the lexicon and phrasing appropriate for articulating their critiques. In Cycle 2, students were coached to author MCQs. By the end of this, each student should have written an MCQ, received peer feedback and edited their questions according to that feedback. In Cycle 3, we determined performance metrics using a low-stakes test and ultimately used that to inform the choice of questions in the W13E. This included student-authored questions from Cycle 2 (SAMCQs) and MCQs derived from three other sources: instructor-authored past paper questions (IAPPQs), textbook-derived questions from Cycle 1 (TDQs), and student-authored, instructor-edited questions



Instructor-authored past paper questions (IAPPQs)

Academically authored and deployed in examinations for several years previous.

Cycle 2: student-authored MCQs (SAMCQs)

Students shown how to author MCQs, how to get inspiration for topics, stems, and options.

Each student:

1. Writes an MCQ
2. Gives feedback to peers
3. Uses feedback to improve question

Cycle 1: textbook-derived questions (TDQs)

Exercises to develop skills of critiquing MCQs and writing actionable feedback.

Each student:

1. Selects a textbook MCQ
2. Critiques and rates submitted MCQs according to Bloom's Taxonomy and syllabus alignment

Top ranked questions checked by academic for alignment to syllabus and correctness.

Cycle 3: trial run

Students answer >40 questions thoughtfully.

Performance metrics used to identify suitable candidates for exam. Questions checked for accuracy and expression only.

Student-authored, instructor-edited questions (SAIEQs)

Questions generated from an activity similar to Cycle 2, and identified by academics at that time to have potential.

Questions modified by academics.

Week 13 exam (W13E)

High-stakes assessment

Mixture of TDQs, IAPPQs, SAMCQs and SAIEQs.

Figure 1. Summary of the process and the origin of the questions in the W13E. Over 600 students, across eight tutorial groups, participated in these activities to find and critique existing MCQs, write their own and test the resultant pool of questions, which also included IAPPQs and SAIEQs. The best performing questions were used for the high-stakes assessment.

(SAIEQs), which had been created in PeerWise exercises the previous year, selected for potential and revised by academics before deployment.

Details of each activity

Cycle 1: Reflecting on the quality of MCQs: learning to identify attributes and articulate opinion

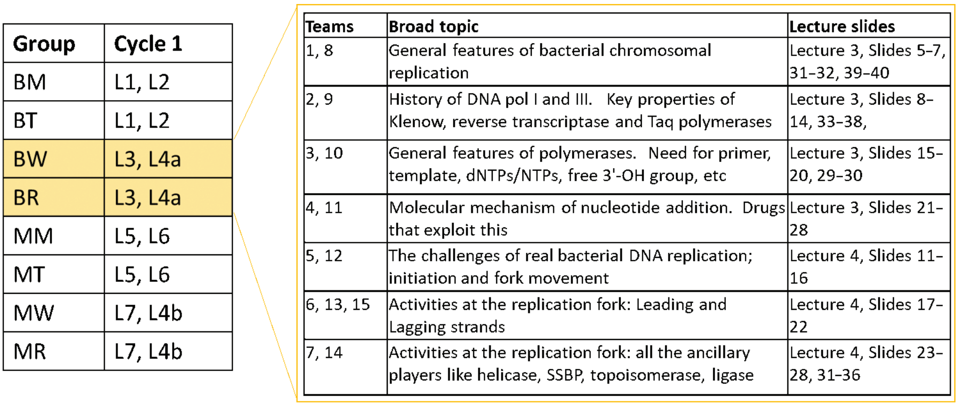
In a tutorial setting, students were introduced to Bloom's Taxonomy and its application to assessment items. Students discussed, in small teams of 5–6, the elements of seven sample MCQs each with four answer options, representing distinct styles and outcomes on the learning achievement spectrum.

The discussions were captured by a scribe on the collaborative whiteboard website Padlet (https://padlet.com/), and the sessions ended with a plenary discussion on those contributions. Students were also encouraged to submit their own reflections to PeerWise, both to gain familiarity with the platform and to practice the language of Bloom's Taxonomy. This activity generated several hundred pieces of feedback on each question and was effective in confirming that students were competent at recognising the key attributes of an MCQ and were able to articulate their opinions.

The assignment for this tutorial was to source a question from a textbook or website on the topic assigned to their team (about 5 lecture slides; Table 1) and to post it for review. Each student had to contribute one question and review 10 other submissions.

Marks were awarded based on the quality of question review. It was also important that questions were submitted on time and appropriately tagged, but question quality was not marked. The differences in question quality were ideal for developing criticism and articulation skills, and critiques could be honest, as they were not authored by the students themselves; 667 questions were submitted (representing a

Table 1. Setup of topic allocations in Cycles 1 and 2.



Group	Cycle 1
BM	L1, L2
BT	L1, L2
BW	L3, L4a
BR	L3, L4a
MM	L5, L6
MT	L5, L6
MW	L7, L4b
MR	L7, L4b

Teams	Broad topic	Lecture slides
1, 8	General features of bacterial chromosomal replication	Lecture 3, Slides 5–7, 31–32, 39–40
2, 9	History of DNA pol I and III. Key properties of Klenow, reverse transcriptase and Taq polymerases	Lecture 3, Slides 8–14, 33–38,
3, 10	General features of polymerases. Need for primer, template, dNTPs/NTPs, free 3'-OH group, etc	Lecture 3, Slides 15–20, 29–30
4, 11	Molecular mechanism of nucleotide addition. Drugs that exploit this	Lecture 3, Slides 21–28
5, 12	The challenges of real bacterial DNA replication; initiation and fork movement	Lecture 4, Slides 11–16
6, 13, 15	Activities at the replication fork: Leading and Lagging strands	Lecture 4, Slides 17–22
7, 14	Activities at the replication fork: all the ancillary players like helicase, SSBP, topoisomerase, ligase	Lecture 4, Slides 23–28, 31–36

There were eight tutorial groups of 70–80 students, which were further split into teams of about six students. For example, BW and BR were allocated the topics shown on the right. This was rotated for Cycle 2 (e.g. BW/BR worked on L5, L6 in Cycle 2).

99% response rate), with 94% of these reviewed by more than 5 students, and 49% reviewed by more than 10 students. Students were also asked to judge each question based on Bloom's Taxonomy using the standard PeerWise rating scale of 0–6 (0 for surface recall, to 6 for creativity/extrapolation) and difficulty scale of 0–2 (0 for easy, to 2 for hard).

After Cycle 1, students received feedback on their critiques and reminded that useful feedback has to be actionable and specific. They were given examples of feedback deemed 'useful' (thoughtful, specific, constructive, insightful, suggestions, expansive, articulate, comment on alignment to syllabus, reference to Bloom's) and 'non-useful' (platitudes, perfunctory, general, repeated aspects of question).

About 40 questions from each of the eight groups were considered for deployment in future activities (i.e. the TDQs). To be eligible for selection, the question had to have at least 10 answers, a single most popular option and a student rating on the Bloom's Taxonomy level of at least two. These questions were then appraised by a subject matter expert (a teaching assistant in the course) to confirm correctness and alignment with the syllabus. This person could reject questions but were prohibited from changing the structure or wording beyond minor edits to improve clarity of expression. This approach to filtering the questions does mean that some potentially good questions were rejected. Questions with fewer than 10 answers, or those with poor student ratings, were not considered for future deployment. The rationale for this was partly to manage the workload of the subject expert for the review phase and also to deliberately include only questions that were discriminatory in the field, rather than relying on instructor judgements of what might be discriminatory, and for this, the questions required a sufficient number of responses.

Cycle 2: Question authoring; development and refinement

This cycle aimed to equip students with the skills to harvest authentic peer misconceptions and profound insights on a specific topic and integrate these into a novel MCQ. Students participated in a tutorial in which they explained concepts to each other, with members of the group challenging, querying and extending these concepts to scope the boundaries of each other's knowledge. Students then used this intelligence to each design a question with four choices on their allocated topic (Table 1).

As with Cycle 1, students were required to give feedback on questions submitted by their peers in line with Bloom's Taxonomy. They were also asked to comment on question structure and the extent to which it prompted reflection of the concept being tested. Student-authors could then use the comments to improve their submission. Only token marks were allocated to reward editing activity, but it was successful at encouraging the revision of 78% of questions. No academic judgement of question quality was made, with marks primarily assigned to comment quality and action based on peer feedback.

Cycle 3: Determination of performance metrics of MCQs

All the questions from Cycle 2 were redeployed in new PeerWise courses to each of the eight classes that produced them (~70 SAMCQs each). We also added ~15 TDQs, 15 IAPPQs and 15 SAIEQs to each of these pools, resulting in a total of around 110 questions for each class or 702 questions in total, as some questions were tested by multiple classes. The SAIEQs were produced in a similar activity to Cycle 2, with the major difference being that they had been specifically assessed by academics and had been tagged as being suitable for further development. Accordingly, these questions were revised by our team of academics to become SAIEQs.

Each of the ~70 students in each class was tasked to answer at least 40 questions from their pool of around 110 questions. Although we were unable to control which questions were answered, >90% of the questions in each pool were answered at least 25 times. As this was a low-stakes task, it was also difficult to ensure that students thoughtfully considered each question. Since the purpose was to obtain accurate performance metrics, it was vital that students did not just choose easy questions or corrupt the data by answering flippantly. Therefore, to reflect an authentic exam setting, students were told that our scoring algorithms would reward genuine attempts, characterised by spending at least 1.5 min per question and submitting defensible quality and difficulty ratings. Students were not required to provide comments, nor were they assessed.

We judged this approach to be successful, as the average time spent by each student was 74.2 ± 3.2 min for the eight groups, which in total submitted over 24 900 answers. A custom dashboard was designed using Filemaker Pro 18 Advanced to easily view the outcomes of student answering activities. Data from students judged to not have taken the task seriously (generally <30 min and undertaken close to the deadline) were omitted from analyses of question performance metrics. Within each of the eight groups, and based on the proportion of questions answered correctly, students were divided into tertiles for the computation of question performance metrics.

The main metrics calculated for each question were difficulty (% correct) and discrimination index (DI; the proportion of the bottom tertile that answered correctly subtracted from the proportion of the top tertile that answered correctly).

Use of intelligence from Cycle 3

Questions with a DI >0.2 and a difficulty of ~60% were considered candidates for inclusion in the high-stakes assessment. This filtered the pool of 702 questions to

~200, including all the IAPPQs, which were automatically included. The performance of the distractors in each of these questions was classified according to two easily implemented heuristics; (1) identification of obvious ambiguities, as shown by students selecting one or two major options, and (2) reflection on the utility of each option, with particular consideration given to the identification of options that were not being picked at all. Although in this iteration this task was performed manually, these attributes can be calculated either for automation or to assist the decision-making process.

Questions with minimal ambiguity and at least three selected distractors were then quality checked by an academic. Only minimal editing was performed; just sufficient to confirm that the question conformed to baseline standards, was aligned to the syllabus, was interpretable, and had a single, genuinely correct, option.

Preparation of high-stakes assessment

Four sets of 42 questions each were prepared for four different versions of the W13E. Due to timetabling constraints, we allowed students to choose any one of the four possible days to complete the 1 h assessment. To mitigate academic integrity risks in an unsupervised exam (primarily collusion and impersonation of a peer), we chose to deliver 42 questions, since we expected this to take the full 1 h based on data from Cycle 3. Furthermore, the order of the questions and answers was shuffled for each student. This assessment contributed 10% to their final mark, and was run in a timed (1 h), online format. Although, from an assessment point of view, we wished to include only questions with ideal DI/difficulty metrics from Cycle 3, we compromised to ensure that each paper had approximately an equal blend of questions from the four sources. The four exams were deployed throughout the week, allowing students to choose which day to sit the task. For fairness, we ensured consistency among question pools in overall difficulty, DIs, expected response time and coverage of learning outcomes.

Determination of question performance

In addition to calculating the difficulty and discrimination factors by traditional methods, we employed two other approaches.

Item response theory (IRT) was used to generate a graphical depiction (item characteristic curves) of how the probability of success within a question was related to student ability. This allows for a more granular and dynamic appraisal of the relationship of these factors than the traditional DI. To generate the characteristic curves for each question, the score obtained by each student was processed in the statistical analysis package, R, according to the workflow outlined by Xie *et al*. (2019) who provide an excellent high-level description of IRT and its differences with the more traditional classical test theory.

Briefly, W13E data were organised in a spreadsheet with each row representing a student and each column representing a question, so that each cell contains a student’s response (1 = correct, 0 = incorrect) to a specific question. This table was submitted to the ‘ltm’ package for R, which displays the probability that students of particular ability would get each question correct. Examples of outputs typical of high- or low-performing questions are shown in the Results.

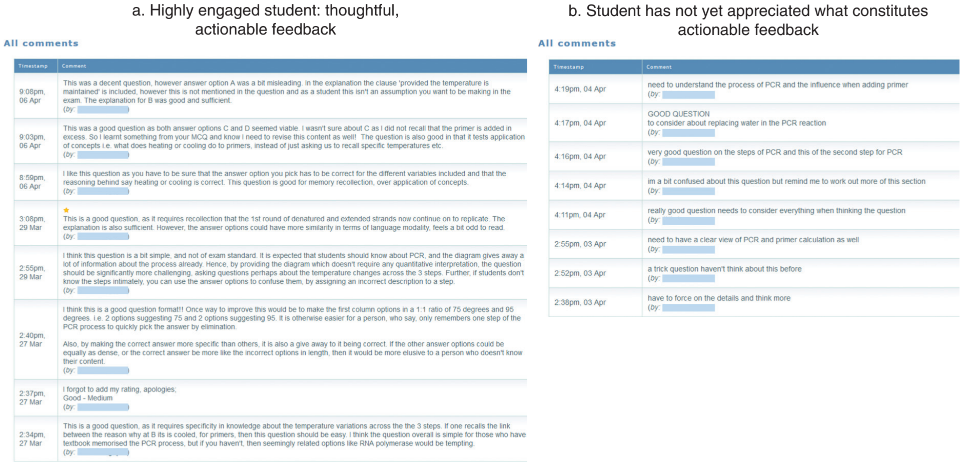
Distractor frequency analysis (DFA) is an in-house developed method for describing the likelihood that cohorts will select particular options within an MCQ. The class was divided into six groups based on overall exam or MCQ component mark, and the percentage of students in each sextile that choose each option was calculated and presented in an easy-to-interpret interface. Examples of outputs typical of high- and low-performing questions are shown in the Results. This type of analysis enabled us to investigate whether certain misconceptions (identified by a high proportion of students selecting a particular distractor) are more prevalent in different student cohorts. If distractors are each designed to test a specific common misconception, then it is beneficial to know which distractors are selected by each of the performance cohorts. For example, if a particular concept is only understood by the top cohort, with other groups primarily selecting one distractor, it would indicate that more attention needs to be given to this concept and the specific misconception targeted by the distractor. If a misconception was more prevalent in only the bottom cohort(s) as in Table 3, then perhaps a different remediation strategy might be used.

Results

Monitoring student contributions

A key component of our strategy to train students to become effective MCQ authors was to encourage them to provide timely, actionable feedback and was therefore our focus in Cycles 1 and 2. PeerWise provides a report that collates all the contributions (questions, comments and replies) from one student on one page. It takes about 30 s to scan the comments and confirm their usefulness or otherwise (Figure 2).

Success of the broad question performance trawl in Cycle 3 depended on students approaching this largely formative task with sincerity, answering each question to the best of their ability. From experience, students generally take a task seriously if they trust that their efforts will be rewarded. Therefore, we developed a dashboard



a. Highly engaged student: thoughtful, actionable feedback

All comments

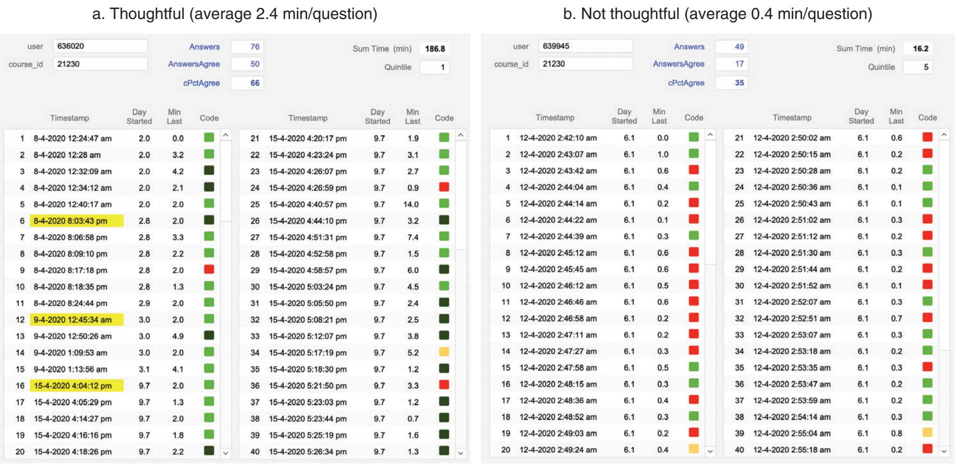
Timestamp	Comment
9:08pm, 06 Apr	This was a decent question, however answer option A was a bit misleading. In the explanation the clause 'provided the temperature is maintained' is included, however this is not mentioned in the question and as a student this isn't an assumption you want to be making in the exam. The explanation for B was good and sufficient. (*by:* [redacted])
9:03pm, 06 Apr	This was a good question as both answer options C and D seemed viable. I wasn't sure about C as I did not recall that the primer is added in excess. So I learnt something from your MCQ and know I need to revise this content as well! The question is also good in that it tests application of concepts i.e. what does heating or cooling do to primers, instead of just asking us to recall specific temperatures etc. (*by:* [redacted])
8:59pm, 06 Apr	I like this question as you have to be sure that the answer option you pick has to be correct for the different variables included and that the reasoning behind say heating or cooling is correct. This question is good for memory recollection, over application of concepts. (*by:* [redacted])
3:08pm, 29 Mar	★ This is a good question, as it requires recollection that the 1st round of denatured and extended strands now continue on to replicate. The explanation is also sufficient. However, the answer options could have more similarity in terms of language modality, feels a bit odd to read. (*by:* [redacted])
2:55pm, 29 Mar	I think this question is a bit simple, and not of exam standard. It is expected that students should know about PCR, and the diagram gives away a lot of information about the process already. Hence, by providing the diagram which doesn't require any quantitative interpretation, the question should be significantly more challenging, asking questions perhaps about the temperature changes across the 3 steps. Further, if students don't know the steps intimately, you can use the answer options to confuse them, by assigning an incorrect description to a step. (*by:* [redacted])
2:40pm, 27 Mar	I think this is a good question format!! Once way to improve this would be to make the first column options in a 1:1 ratio of 75 degrees and 95 degrees. i.e. 2 options suggesting 75 and 2 options suggesting 95. It is otherwise easier for a person, who say, only remembers one step of the PCR process to quickly pick the answer by elimination. Also, by making the correct answer more specific than others, it is also a give away to it being correct. If the other answer options could be equally as dense, or the correct answer be more like the incorrect options in length, then it would be more elusive to a person who doesn't know their content. (*by:* [redacted])
2:37pm, 27 Mar	I forgot to add my rating, apologies; Good - Medium (*by:* [redacted])
2:34pm, 27 Mar	This is a good question, as it requires specificity in knowledge about the temperature variations across the the 3 steps. If one recalls the link between the reason why at B its is cooled, for primers, then this question should be easy. I think the question overall is simple for those who have textbook memorised the PCR process, but if you haven't, then seemingly related options like RNA polymerase would be tempting. (*by:* [redacted])

b. Student has not yet appreciated what constitutes actionable feedback

All comments

Timestamp	Comment
4:19pm, 04 Apr	need to understand the process of PCR and the influence when adding primer (*by:* [redacted])
4:17pm, 04 Apr	GOOD QUESTION to consider about replacing water in the PCR reaction (*by:* [redacted])
4:16pm, 04 Apr	very good question on the steps of PCR and this of the second step for PCR (*by:* [redacted])
4:14pm, 04 Apr	im a bit confused about this question but remind me to work out more of this section (*by:* [redacted])
4:11pm, 04 Apr	really good question needs to consider everything when thinking the question (*by:* [redacted])
2:55pm, 03 Apr	need to have a clear view of PCR and primer calculation as well (*by:* [redacted])
2:52pm, 03 Apr	a trick question haven't think about this before (*by:* [redacted])
2:38pm, 03 Apr	have to force on the details and think more (*by:* [redacted])

Figure 2. Comments of an engaged and not so engaged student to the same questions in Cycle 2. The feedback from Student A can be used to improve the question. Moreover, they are more likely to reflect more on the attributes of their own question. Outputs like this were valuable in showing students how they should develop their articulation skills.



a. Thoughtful (average 2.4 min/question)

user 636020
course_id 21230
Answers 76
AnswersAgree 50
cPctAgree 66
Sum Time (min) 186.8
Quintile 1

	Timestamp	Day Started	Min Last	Code
1	8-4-2020 12:24:47 am	2.0	0.0	🟩
2	8-4-2020 12:28 am	2.0	3.2	🟩
3	8-4-2020 12:32:09 am	2.0	4.2	■
4	8-4-2020 12:34:12 am	2.0	2.1	■
5	8-4-2020 12:40:17 am	2.0	2.0	🟩
6	8-4-2020 8:03:43 pm	2.8	2.0	■
7	8-4-2020 8:06:58 pm	2.8	3.3	🟩
8	8-4-2020 8:09:10 pm	2.8	2.2	🟩
9	8-4-2020 8:17:18 pm	2.8	2.0	🟥
10	8-4-2020 8:18:35 pm	2.8	1.3	🟩
11	8-4-2020 8:24:44 pm	2.9	2.0	🟩
12	9-4-2020 12:45:34 am	3.0	2.0	🟩
13	9-4-2020 12:50:26 am	3.0	4.9	■
14	9-4-2020 1:09:53 am	3.0	2.0	🟩
15	9-4-2020 1:13:56 am	3.1	4.1	🟩
16	15-4-2020 4:04:12 pm	9.7	2.0	🟩
17	15-4-2020 4:05:29 pm	9.7	1.3	🟩
18	15-4-2020 4:14:27 pm	9.7	2.0	🟩
19	15-4-2020 4:16:16 pm	9.7	1.8	🟩
20	15-4-2020 4:18:26 pm	9.7	2.2	■

	Timestamp	Day Started	Min Last	Code
21	15-4-2020 4:20:17 pm	9.7	1.9	🟩
22	15-4-2020 4:23:24 pm	9.7	3.1	🟩
23	15-4-2020 4:26:07 pm	9.7	2.7	🟩
24	15-4-2020 4:26:59 pm	9.7	0.9	🟥
25	15-4-2020 4:40:57 pm	9.7	14.0	🟩
26	15-4-2020 4:44:10 pm	9.7	3.2	■
27	15-4-2020 4:51:31 pm	9.7	7.4	🟩
28	15-4-2020 4:52:58 pm	9.7	1.5	🟩
29	15-4-2020 4:58:57 pm	9.7	6.0	■
30	15-4-2020 5:03:24 pm	9.7	4.5	🟩
31	15-4-2020 5:05:50 pm	9.7	2.4	■
32	15-4-2020 5:08:21 pm	9.7	2.5	■
33	15-4-2020 5:12:07 pm	9.7	3.8	■
34	15-4-2020 5:17:19 pm	9.7	5.2	🟨
35	15-4-2020 5:18:30 pm	9.7	1.2	■
36	15-4-2020 5:21:50 pm	9.7	3.3	🟥
37	15-4-2020 5:23:03 pm	9.7	1.2	■
38	15-4-2020 5:23:44 pm	9.7	0.7	■
39	15-4-2020 5:25:19 pm	9.7	1.6	■
40	15-4-2020 5:26:34 pm	9.7	1.3	■

b. Not thoughtful (average 0.4 min/question)

user 639945
course_id 21230
Answers 49
AnswersAgree 17
cPctAgree 35
Sum Time (min) 16.2
Quintile 5

	Timestamp	Day Started	Min Last	Code
1	12-4-2020 2:42:10 am	6.1	0.0	🟩
2	12-4-2020 2:43:07 am	6.1	1.0	🟩
3	12-4-2020 2:43:42 am	6.1	0.6	🟥
4	12-4-2020 2:44:04 am	6.1	0.4	🟩
5	12-4-2020 2:44:14 am	6.1	0.2	🟥
6	12-4-2020 2:44:22 am	6.1	0.1	🟥
7	12-4-2020 2:44:39 am	6.1	0.3	🟩
8	12-4-2020 2:45:12 am	6.1	0.6	🟥
9	12-4-2020 2:45:45 am	6.1	0.6	🟥
10	12-4-2020 2:46:12 am	6.1	0.5	🟥
11	12-4-2020 2:46:46 am	6.1	0.6	🟥
12	12-4-2020 2:46:58 am	6.1	0.2	🟥
13	12-4-2020 2:47:11 am	6.1	0.2	🟥
14	12-4-2020 2:47:27 am	6.1	0.3	🟥
15	12-4-2020 2:47:58 am	6.1	0.5	🟩
16	12-4-2020 2:48:15 am	6.1	0.3	🟩
17	12-4-2020 2:48:36 am	6.1	0.4	🟥
18	12-4-2020 2:48:52 am	6.1	0.3	🟩
19	12-4-2020 2:49:03 am	6.1	0.2	🟥
20	12-4-2020 2:49:24 am	6.1	0.4	🟨

	Timestamp	Day Started	Min Last	Code
21	12-4-2020 2:50:02 am	6.1	0.6	🟥
22	12-4-2020 2:50:15 am	6.1	0.2	🟥
23	12-4-2020 2:50:28 am	6.1	0.2	🟩
24	12-4-2020 2:50:36 am	6.1	0.1	🟩
25	12-4-2020 2:50:43 am	6.1	0.1	🟩
26	12-4-2020 2:51:02 am	6.1	0.3	🟥
27	12-4-2020 2:51:12 am	6.1	0.2	🟥
28	12-4-2020 2:51:30 am	6.1	0.3	🟩
29	12-4-2020 2:51:44 am	6.1	0.2	🟥
30	12-4-2020 2:51:52 am	6.1	0.1	🟥
31	12-4-2020 2:52:07 am	6.1	0.3	🟩
32	12-4-2020 2:52:51 am	6.1	0.7	🟥
33	12-4-2020 2:53:07 am	6.1	0.3	🟩
34	12-4-2020 2:53:18 am	6.1	0.2	🟩
35	12-4-2020 2:53:35 am	6.1	0.3	🟥
36	12-4-2020 2:53:47 am	6.1	0.2	🟩
37	12-4-2020 2:53:59 am	6.1	0.2	🟩
38	12-4-2020 2:54:14 am	6.1	0.3	🟩
39	12-4-2020 2:55:04 am	6.1	0.8	🟨
40	12-4-2020 2:55:18 am	6.1	0.2	🟥

Figure 3. Custom dashboard developed to monitor student engagement in Cycle 3. Answering patterns are shown, including successful attempts on difficult questions (dark green), unsuccessful attempts on easy questions (red) and commencement of any sessions following the first one (yellow highlights). Data from students such as Student B were excluded from analyses of question performance.

Table 2. Statistics of the four W13E.

Day	Students	Questions	High score (%)	Low score (%)	Average score (%)	SD	Average time (min:s)
Monday	39	42	93	38	72	6.52	56:55
Tuesday	68	42	98	40	75	5.99	57:47
Wednesday	169	42	98	38	73	5.81	58:07
Thursday	325	42	95	21	67	6.24	57:55

(Figure 3) using granular time-stamped activity data (available on request from PeerWise), to convince students that we could see their approach to the questions. This strategy proved successful as the vast majority (>87%) of students completed the task with appropriate diligence, judged only from inspection of the start and end times of answering sessions. Such students spent 1–2 h answering the 40 questions, and, reassuringly, 12% of students spent less than 1 min per question. Only 5% were judged to have not performed the task thoughtfully, as evidenced by them spending less than 30 s per question, selecting mainly 'easy' questions, and still answering the majority incorrectly.

Overall exam information

Over 600 students sat one of the four W13E. Students could choose any 1 day to sit the exam, with nearly 55% of them taking the final, Thursday paper. Despite each exam consisting of largely different questions, with only 2–3 questions being redeployed on other days, each paper was of approximately equal difficulty. Both the average and lowest exam scores were reduced for the Thursday version (Table 2), but the cohort of students who took this exam were slightly weaker, as indicated by

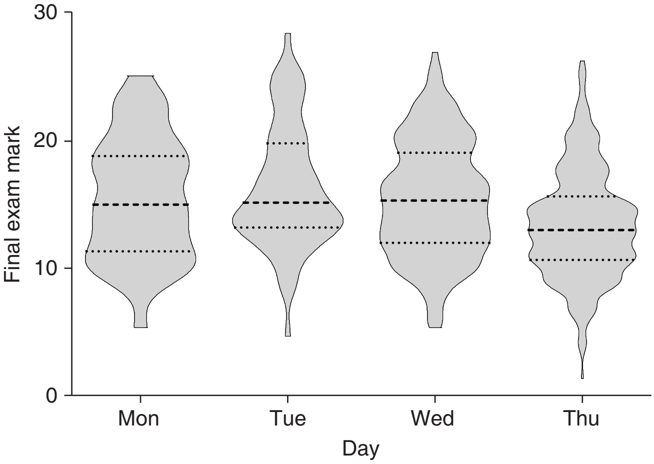


Figure 4. Performance of students in the final exam. The marks of the end-of-semester exam grouped by the day on which the students sat the W13E. The dashed and dotted lines represent the median and quartiles, respectively.

their performance in the formal end-of-semester examination (Figure 4). We did not include questions from the Monday exam in further analyses, due to the low number of students.

There was a strong linear relationship between the W13E (containing >50% student-authored questions) and the marks of the Final Exam (all academic written questions) with the correlation being particularly strong for the higher-performing students (Figure 5).

Question performance analysis and classification

The performance of each of the 126 questions from the Tuesday, Wednesday and Thursday exams was assessed according to three separate analytical techniques: Classical, IRT and DFA. In the IRT, the ability of the students is plotted against the probability that they will answer the question correctly. An ability of 0 represents students performing at an average standard, and 4 and −4 representing the highest and lowest achieving students across the entire exam.

In the example (Table 3), the probability that a student of average ability will get this question correct is over 0.6, and over 95% of high-achieving students and less than 5% of students at the bottom end are getting it correct. The strong performance of this question is supported by the Classical analysis, which shows that 64% of the class chose the correct option (B) and that the DI for this option (and, therefore the question as a whole) was 0.33, for example, the proportion of students choosing option B was much higher in the stronger students than in the bottom sextile. Note that the false option, A, shows negative DI, which is desirable.

DFA expands on the classical approach by showing the proportion of each sextile that chooses a particular distractor. Q0 is the top sextile and Q5 is the bottom. Correct choices selected by more than 50% of the sextile are highlighted in pink and wrong choices selected by more than 25% are in blue. This complements the other analyses by providing intelligence on the likelihood that particular cohorts will select specific

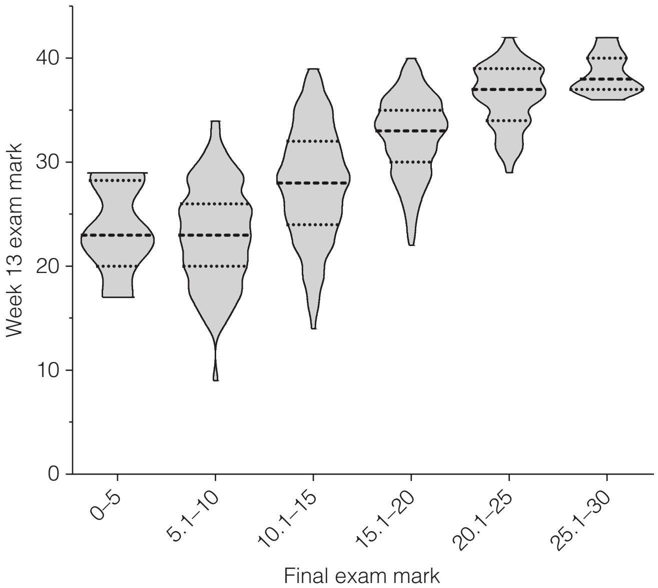
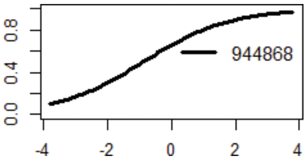


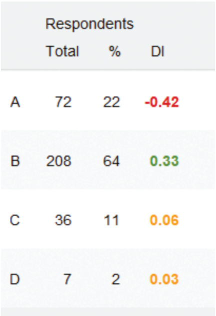
Figure 5. Comparison of student performance in the W13E and end-of-semester exam. Marks from the W13E were plotted for each student, grouped by their performance in the end-of-semester exam. The dashed and dotted lines represent the median and quartiles, respectively.

Table 3. Performance metrics for 'Good – Standard' question.

IRT

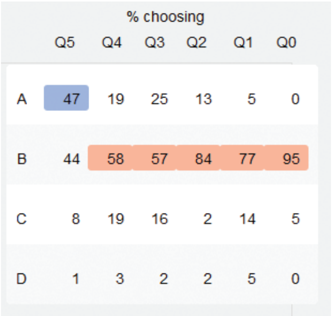


Classical



	Respondents		
	Total	%	DI
A	72	22	-0.42
B	208	64	0.33
C	36	11	0.06
D	7	2	0.03

DFA



	% choosing					
	Q5	Q4	Q3	Q2	Q1	Q0
A	47	19	25	13	5	0
B	44	58	57	84	77	95
C	8	19	16	2	14	5
D	1	3	2	2	5	0

The x-axis of the IRT analysis represents ability, and the y-axis represents probability that they answer correctly. This appeared in the Thursday paper, in a random position for each student. The number 944868 is the unique ID of the question allocated within the online delivery platform. The DFA shows the % students who chose each option (A, B, C or D) for each of the 6 cohorts (divided up by performance in the overall exam); Q5 being the lowest performing cohort through to Q0 being the top-performing students. The pink shading indicates the correct option while the blue shading shows a popular distractor, in this case only in the lowest cohort.

options, and if the options appear ambiguous to particular groups of students, which may or may not be desirable.

Using performance metrics, we classified each question into five main categories: Good – Barrier (deliberately easy questions that all moderately engaged students should answer correctly), Good – Standard (discriminating effectively between passing and higher achieving students), Good – Difficult (discriminating between students at the top of the class), Poor – Noise (answered effectively randomly by students of all abilities) and Poor – Negative (weaker students more likely to select the 'correct' answer, actively penalises the higher achieving students) (Table 4). Categorisation of each question was agreed on by each researcher with little ambiguity.

For very easy questions, the IRT curve is flat and consistently high, with students at both top and bottom sextiles answering this question correctly, and with none of the distractors being attractive to any particular cohort. It is arguable as to whether a question of this nature should be classified as 'good' or 'bad'. While it does not discriminate between high- and low-achieving cohorts of students, if pitched to confirm achievement of a core learning outcome, it is a worthy measure of student achievement. Academics are increasingly encouraged to include barrier questions such as this, but only 10 of the 126 questions in our exams were so classified.

The majority of questions (76/126) in our exams were classified as a good standard question, which discriminated between passing and failing students. The IRT curve shows that students in the higher ranges of ability all answer this question correctly, and even those of medium ability get it correct more than 50% of the time, whereas very poor students do not get it correct. This is supported by the traditional metrics, which show a correctness/difficulty of 61% and a DI of 0.47. Further reinforcement of the classification of this question comes from the DFA, which confirms that over 90% of the top sextile and around 40% of the fourth sextile (representing the passing students) answered it correctly. However, the extra intelligence obtained from DFA is that Option A was particularly attractive to weaker students and Option D was a very weak distractor for all cohorts.

Questions were classified as hard if they discriminated between students towards the top of the class. The IRT curve is shifted to the right, the overall correctness is reduced to 30% and the DI is >0.3. In addition, the DFA confirms that 80% of the top sextile got the question correct, with every other sextile favouring Option C except the weakest students who opted more frequently for Option D.

One indication of a poor question is a flat IRT curve in which the question is answered effectively in a random manner across all abilities. Indeed, DFA and Classical analyses confirm that students were reduced to taking a 50:50 bet between two obviously ambiguous options. A question like this generates noise in the student scores but does not have an overtly negative effect on either cohort.

The worst type of question is revealed in the last row of the table. The IRT plot has a negative slope, the DI is negative and Classical analysis shows that students were choosing between two options. However, the better the student, the more likely they were to select the option that had been classified as incorrect. None of the 126 questions used in our analyses were classified in this way.

Analysis by traditional metrics

Based on traditional metrics, the SAMCQs and SAIEQs were at least as good as academic-authored questions, whether they were past paper questions or those sourced from the textbook or web (Figure 6a). Indeed, contrary to the latter two groups of questions, the DIs of all SAIEQs were positive. Questions from the textbook/web contained several questions with low or negative DIs, and the greatest spread in the DI range was seen in the past paper questions.

Table 4. Summary of the types of questions that appeared in the exam.

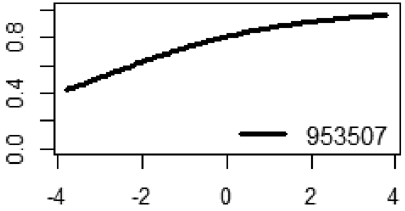
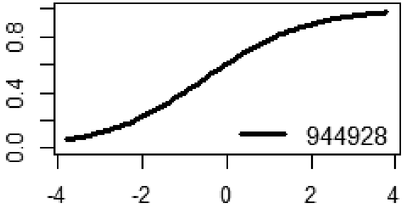
Question	IRT	Classical	DFA
Barrier *n* = 10		Respondents Total % DI A 21 7 0.00 B 15 5 -0.06 C 259 81 0.11 D 26 8 -0.06	% choosing S5 S4 S3 S2 S1 S1+ A 4 7 11 8 5 0 B 8 7 2 3 2 0 C 75 72 82 85 86 100 D 12 13 5 3 7 0
Standard *n* = 76		Respondents Total % DI A 91 29 -0.32 B 192 61 0.47 C 30 9 -0.11 D 4 1 -0.04	% choosing S5 S4 S3 S2 S1 S1+ A 41 42 29 21 9 5 B 44 42 51 77 91 95 C 11 15 20 2 0 0 D 4 1 0 0 0 0

Table 4. (Continued)

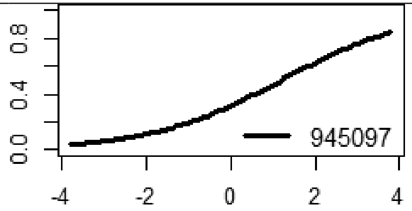
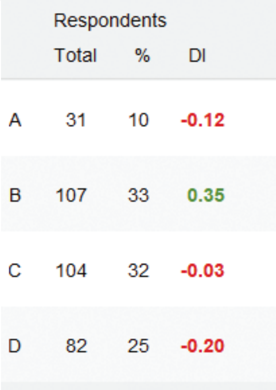
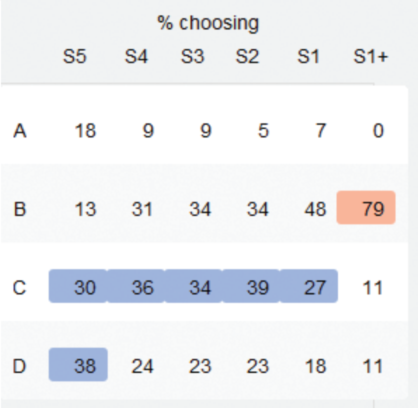
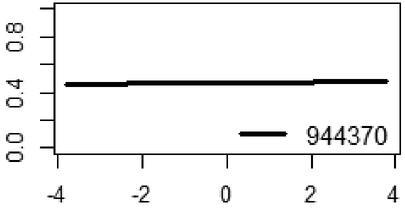
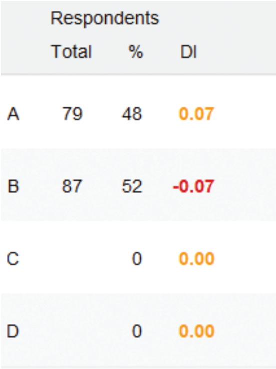
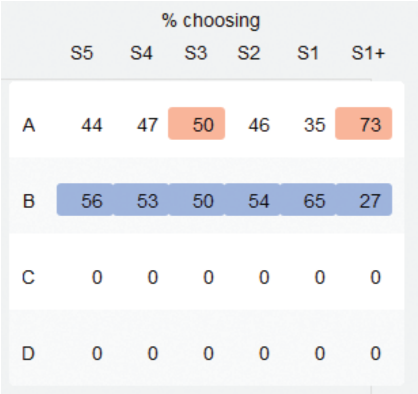
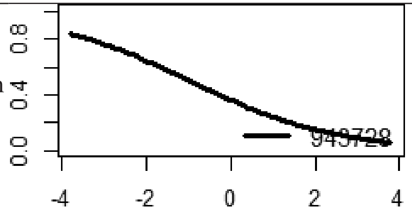
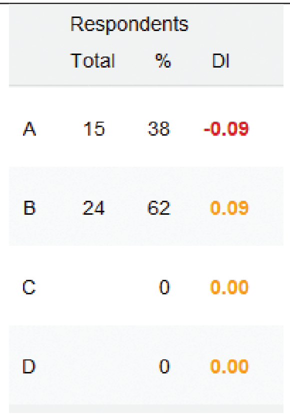
<table>
<tr><th>Question</th><th>IRT</th><th>Classical</th><th>DFA</th></tr>
<tr><td>Hard
<i>n</i> = 36</td><td> </td><td>
<table>
<tr><th></th><th colspan="3">Respondents</th></tr>
<tr><th></th><th>Total</th><th>%</th><th>DI</th></tr>
<tr><td>A</td><td>31</td><td>10</td><td>-0.12</td></tr>
<tr><td>B</td><td>107</td><td>33</td><td>0.35</td></tr>
<tr><td>C</td><td>104</td><td>32</td><td>-0.03</td></tr>
<tr><td>D</td><td>82</td><td>25</td><td>-0.20</td></tr>
</table></td><td>
<table>
<tr><th></th><th colspan="6">% choosing</th></tr>
<tr><th></th><th>S5</th><th>S4</th><th>S3</th><th>S2</th><th>S1</th><th>S1+</th></tr>
<tr><td>A</td><td>18</td><td>9</td><td>9</td><td>5</td><td>7</td><td>0</td></tr>
<tr><td>B</td><td>13</td><td>31</td><td>34</td><td>34</td><td>48</td><td>79</td></tr>
<tr><td>C</td><td>30</td><td>36</td><td>34</td><td>39</td><td>27</td><td>11</td></tr>
<tr><td>D</td><td>38</td><td>24</td><td>23</td><td>23</td><td>18</td><td>11</td></tr>
</table></td></tr>
<tr><td>Noise
<i>n</i> = 4</td><td> </td><td>
<table>
<tr><th></th><th colspan="3">Respondents</th></tr>
<tr><th></th><th>Total</th><th>%</th><th>DI</th></tr>
<tr><td>A</td><td>79</td><td>48</td><td>0.07</td></tr>
<tr><td>B</td><td>87</td><td>52</td><td>-0.07</td></tr>
<tr><td>C</td><td></td><td>0</td><td>0.00</td></tr>
<tr><td>D</td><td></td><td>0</td><td>0.00</td></tr>
</table></td><td>
<table>
<tr><th></th><th colspan="6">% choosing</th></tr>
<tr><th></th><th>S5</th><th>S4</th><th>S3</th><th>S2</th><th>S1</th><th>S1+</th></tr>
<tr><td>A</td><td>44</td><td>47</td><td>50</td><td>46</td><td>35</td><td>73</td></tr>
<tr><td>B</td><td>56</td><td>53</td><td>50</td><td>54</td><td>65</td><td>27</td></tr>
<tr><td>C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>D</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
</table>

Table 4. (Continued)

<table>
<tr><th>Question</th><th>IRT</th><th>Classical</th><th>DFA</th></tr>
<tr>
<td>Negative
$n = 0$
(example from Mon exam)</td>
<td>
</td>
<td>
<table>
<tr><th></th><th colspan="3">Respondents</th></tr>
<tr><th></th><th>Total</th><th>%</th><th>DI</th></tr>
<tr><td>A</td><td>15</td><td>38</td><td>-0.09</td></tr>
<tr><td>B</td><td>24</td><td>62</td><td>0.09</td></tr>
<tr><td>C</td><td></td><td>0</td><td>0.00</td></tr>
<tr><td>D</td><td></td><td>0</td><td>0.00</td></tr>
</table></td>
<td>Too few respondents to be meaningful</td>
</tr>
</table>

The *x*-axis of the IRT analysis represents ability, and the *y*-axis represents probability that they answer correctly. For the DFA shown here, the students have been divided into sextiles based on performance in the overall exam with S5 the poorest performing cohort through to S1+ the top performing students. Again, the pink shading highlights the correct option and the blue shading popular distractors selected by some cohorts.

Textbook and web-sourced questions were also overall easier than student-authored and past paper questions (Figure 6b). As with DIs, the past paper questions also had a greater range of difficulties than the student-authored questions. Student-authored questions with and without contribution by an academic exhibited a smaller range and were of similar difficulty to the harder past paper questions.

Internal calibration of past paper questions

The performance of the IAPPQs was particularly important to define as this serves as an internal control, allowing us to determine if (1) the class were of a similar standard to previous cohorts and (2) we were getting the students to engage with the material in a similar way despite some changes in lecturing staff. This was particularly important, since the 2019 questions were part of a larger, traditionally administered examination that covered the entire syllabus including both molecular biology and metabolism concepts (15 lectures of each). In contrast, the 2020 questions were part of a smaller, shorter, online MCQ exam that covered only half the molecular biology material (7 lectures). Another difference between the 2 years was that the 2020 students had the option to choose which of the four exams they sat.

IAPPQ8 was typical of most of the easier IAPPQs, with students towards the lower end of the class getting this right more than 50% of the time and with identical DIs and derivative curves (Table 5). IAPPQ11 performed similarly between the 2020 and 2019 exams, even though it was a harder question (as illustrated by the peak in the derivative curves at higher ability students), despite it being run on two separate exams in 2020. IAPPQ6 behaved slightly differently between the 2 years, with the 2020 deployment being a better discriminator largely because the less able students answered it more poorly. Overall, 12 IAPPQs were deployed in both 2019 and 2020, and in every case, the performance of each was consistent with the examples shown in the table.

The peaks of the derivative curves are the inflexion points of each IRT curve, for example, the ability level that gives a 50% probability of getting the question right. The spread indicates the extent to which the question discriminates students around this point. The curves were obtained by plotting the first derivative of the IRT plot equation against the ability level (−4 to 4). Sharp peaks are items with strong

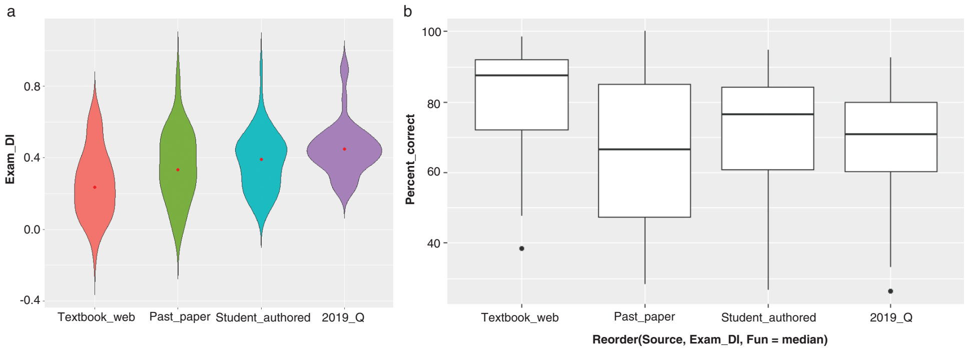


Figure 6. Comparison of traditional metrics of questions from the four sources. The sources from left to right: TDQ, IAPPQ, SAMCQ, SAIEQ, and their (a) discrimination indices and (b) difficulty.

Table 5. Comparison of performance of past paper questions.

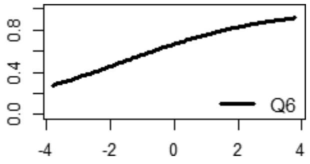
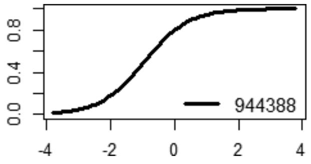
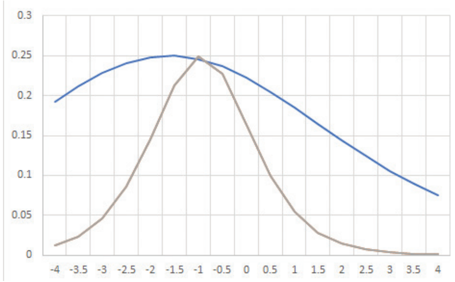
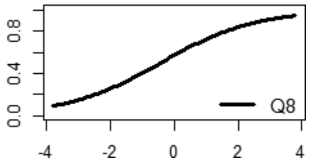
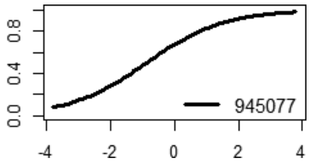
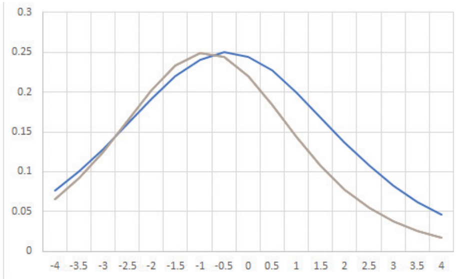
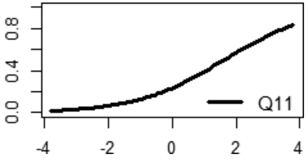
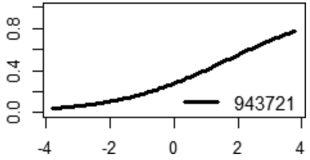
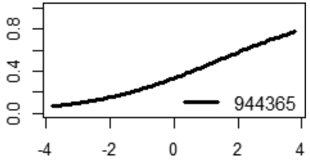
Question	2019 exam	2020 exam	Derivative curves
IAPPQ6	*n* = 719 DI = 0.27 	*n* = 169 DI = 0.55 	Maximum amplitude ~1 
IAPPQ8	*n* = 719 DI = 0.44 	*n* = 325 DI = 0.41 	Maximum amplitude at ability ~−1 

Table 5. (Continued)

Question	2019 exam	2020 exam	Derivative curves
IAPPQ11	*n* = 719 DI = 0.41 	*n* = 68 DI = 0.33  *n* = 325 DI = 0.33 	Maximum amplitude at ability ~1.5

The *x*-axis of the IRT analysis represents ability, and the *y*-axis represents probability that they answer correctly.

Table 6. Quality of questions from the four sources.

	TDQ		IAPPQ		SAMCQ		SAEIQ	
Good								
- Easy	19	(63%)	11	(55%)	27	(59%)	19	(63%)
- Hard	4	(13%)	7	(35%)	14	(30%)	11	(37%)
- Barrier	5	(17%)	2	(10%)	3	(7%)	0	(0%)
Poor	2	(7%)	0	(0%)	2	(4%)	0	(0%)
Total	30		20		46		30	

discrimination between the top and bottom of the cohort, while broader peaks show less discrimination.

Using the 2-parameter logistic model, the probability of student *i* correctly answering question *j* is given by:

$$P_j\left(\theta_i\right)=\frac{e^x}{1+e^x}$$

where $x=a_j\left(\theta_i-b_j\right)$, *a* is the discrimination of the question*, b* is the difficulty of the question and θ is the ability of the student. Thus, a student has a 0.5 probability of answering a question correctly if student ability equals the difficulty of the question.

The first derivative of the equation above is:

$$\frac{e^x}{\left(1+e^x\right)^2}$$

Using the difficulty and discrimination coefficients calculated by the R model program, it was thus possible to calculate and plot ability levels between −4 and 4 using this equation to produce the first derivative plots.

Analysis by IRT and DFA classification

All of the student-authored questions that had been edited by an academic were classified as good and contained both easy and hard questions. The past paper questions also performed similarly, with the addition of barrier questions. Out of the four questions that were categorised as poor, two of these questions were textbook and web-sourced questions, and two were student-authored questions that had only been minimally edited (Table 6).

Indeed, out of the 126 questions analysed, there were only three questions with particularly low DIs; two of these questions were sourced from the textbook/web, with the other being a student-authored question that had not been edited by an academic.

Comparison of question performance in low-stakes versus high-stakes tests

The low-stakes formative activity in Cycle 3 was an important part of the process of selecting suitable questions for the high-stakes summative W13E. Presumably, students were taking more care when completing the summative assessment than the formative activity. To determine the extent to which the metrics could predict the suitability of questions in a real exam, we compared DIs (computed on the basis of tertiles) and difficulties of the questions in the two exams (Figure 7). There was broad agreement in question performance between the exams, particularly in the DI range 0.2–0.6. However, there was a greater spread at very high and low exam DI levels. This is likely because the Cycle 3 DI values are based on fewer responses and a less precise scale of esteem, and are therefore less reliable. As expected, a greater proportion of students answered correctly in the exam than in the low-stakes assessment.

Discussion

In this study, we describe a scalable approach for training students to author high-quality MCQs, and we compare two approaches for selecting student-authored questions to appear on summative assessments. We found that questions selected by either approach performed as effectively as academic-authored past paper and textbook-sourced MCQs.

Several studies have reported crowd-sourcing questions from students, not only to build up revision banks but also as a form of revision itself, since the process of authoring and critiquing questions has been shown to raise student performance. However, because high-stakes summative assessments must be error-free, student-generated questions are rarely used directly for this purpose (Schullo-Feulner *et al.* 2014). This seems to be the case even when the questions have been subject to quite stringent academic review. For example, Harris *et al.* (2015) described an activity where student-authored questions were sequentially reviewed by their peers and experts for factual accuracy and distractor quality, with approved questions published for test practice. However, the questions were not deployed in high-stakes summative exams. Amini *et al.* went a step further, combining approved student-authored questions with teacher-authored questions in a medical imaging exam for radiology students (Amini *et al.* 2020). Analysis of the exam responses revealed that student-authored questions were easier, had significantly more non-functional distractors and were more likely to rely on recall skills compared to teacher-constructed questions. However, they did not report any explicit training or coaching of students in question authoring and, unlike us, did not actively select student-authored questions with a high DI.

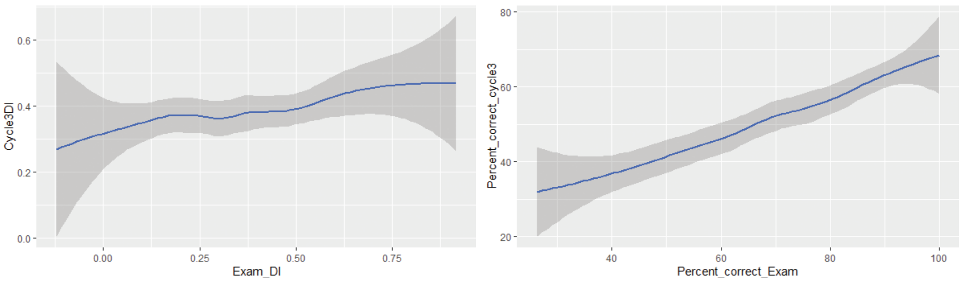


Figure 7. Comparison of question performance in a low-stakes and a high-stakes assessment. The (a) discrimination indices and (b) difficulty of the W13E were plotted against metrics calculated from Cycle 3.

Two approaches for selecting student-authored questions

We examined two distinct approaches for selecting student-authored questions for deployment on a high-stakes summative exam:

1. Identifying candidates by trialling the questions in low-stakes quizzes, followed by light editing to ensure compliance with learning outcomes and basic delivery standards.
2. Identifying potential questions using manual academic judgement followed by intense editing by subject matter experts.

Both methods delivered suitable questions but each comes with quite different overheads in terms of workload and practicability. Filtering and candidate selection using a list of automatically generated performance metrics from low-stakes quizzes are much easier and faster than academic selection involving manual assessment of the stem, options, structure, logic and expression of individual questions.

A key advantage of the low-stakes trial approach was that it provided a reasonable indication of how the questions were likely to perform in the summative exam, something that is difficult to predict from untrialled questions, even for experienced academics. In addition, the selection criteria can be varied to deliver barrier-style (high easiness) or higher-level MCQs as required. Additionally, selection is not biased in the structure or sentiment of the question. In contrast, each academic will have a bias for/against certain MCQ structures and formats. Questions written by students may be dissonant to an academic's normal style, and it can be unsettling to run an assessment using these questions. The increased richness in the range of question styles produced by students provides an opportunity to new ways of thinking and is strongly student-centred, by definition arising from their own stylistic preferences. When faced with the manual selection approach, we noted that as academics, we tended to choose questions that show creativity, insight, extrapolation and effort. This leads to the selection of items that are at the higher levels of Bloom's Taxonomy. However, more complex questions require more work to ensure it is not ambiguous and that each option functions as a suitable distractor. The reward is potentially high because there is a strong chance that the process will produce questions that are innovative and discriminatory and are also strongly aligned to the values, objectives and style of the individual academic, but the time investment is substantial.

A caveat to the low-stakes trial approach is that the metrics can be deceptive. Poor questions can have a high DI if high-achieving students are gravitating to the least incorrect option. Conversely, potentially excellent questions can have a poor (or even negative) DI because strong students are choosing incorrect options that are unattractive to average ability students. For this reason, selection solely based on metrics is not recommended, but it is an efficient initial filter. Similarly, academic judgement alone is not always a reliable predictor of question performance. For example, SAIEQs that did not perform well in Cycle 3 were not included in the W13E. It is likely that, with the benefit of reflection on the performance metrics, some extra revision of the questions would have provided sufficient remediation.

Regardless of the question selection process, using published student-generated questions on high-stakes examinations introduces a concern that students who either authored or answered the questions prior to the exam may have an unfair advantage. There is evidence that students perform significantly better on subsequent exam questions if they have previously authored or answered questions that merely target the same topics (Kelley *et al*. 2019). Similar research, which addresses a potential bias relating to topic-selection, has shown that these results are robust even when students are randomly assigned topics on which to author questions (Denny *et al*. 2017). This concern is partially mitigated in our study as there was a very large question pool and all students had exposure to the entire repository before selection. One potential solution to eliminate any direct advantage would be to stagger the authoring and filtering tasks between academic years or cohorts of students.

Sourcing questions from textbooks

Intuitively, textbook-derived questions (TDQs) would seem to provide the most efficient option, since each question has, presumably, been crafted to reflect fundamental concepts and may even have been tested in genuine assessments. However, even though we chose our 30 TDQs from a pool of some 600 candidates, guided by data from student ratings and tempered with academic judgement, with reference to our specific learning outcomes, these performed no better than any other type of question. Our perception that TDQs are mainly useful for assessing generic facts was supported by the fact that 80% were classified as 'easy' or 'barrier', with just 13% being 'hard'. Student ratings in Cycle 1 rarely rated the TDQs above the most basic levels of Bloom's Taxonomy. Another issue with TDQs is that most academics present their subject matter with their own distinctive emphasis, and so, it is often difficult to make 'generic' TDQs feel relevant to a bespoke syllabus.

Sourcing questions from academics

Whilst it was not practical to obtain an objective measure of the workload associated with the construction of IAPPQs, because these already existed, by definition, we are confident that our introductory assertion – that good quality, field-tested questions are precious – would resonate with most academics.

It would be inappropriate to suggest that all academic-authored questions are as strong as our IAPPQs, which have been constructed by our most student-centred colleagues and fine-tuned over many years of deployment. Indeed, it would not be uncommon for course coordinators to feel some disappointment with questions received from academic colleagues. MCQs from less student-engaged academics can often be hastily constructed and without sufficient insight in the abilities of the student cohort. It is our perception that some research-focused academics favour writing questions that are pitched towards the high-achieving students, and many do not review the performance of their questions, nor refine them from year to year. Of course, novel academic-created MCQs all suffer from not being field tested or subjected to the same cyclical process of review, discussion and editing that occurs in a tool like PeerWise.

We would argue that academics could also benefit from classes on the strategic approach to MCQ authoring. It is sometimes hard for us, as expert practitioners, to see the topic through the lens of a developing learner. Even when given a very specific concept on which to design an MCQ, it can be difficult to know what facet of that topic should be assessed and the academic level at which the question should be pitched. Conversely, when our understanding is developing, it is the moments of enlightenment that define our advancing mastery of the topic, and it is these that form the basis of the most meaningful correct options in MCQs.

Training and associated workload

We adopted PeerWise in the current study, as it is free to use, has excellent in-built reports on student activity, provides low-level access to data for the construction of bespoke dashboards and has been a fixture in our course for several years. The benefits to student academic performance result from all aspects of engaging with PeerWise: answering questions, articulating criticism of peers' questions, authoring questions and editing questions in light of peer feedback (Doyle & Buckley 2020; Hancock *et al*. 2018; Hardy *et al*. 2014; Hudson, Jarstfer & Persky 2018; Kay, Hardy & Galloway 2018; Kay, Hardy & Galloway 2020; Walsh *et al*. 2018). When students engage in discussions that reveal insights and misconceptions, the reflective and critical processes involved are likely to contribute to the learning experience. In our study, we put an emphasis on training students to create and criticise MCQs through scaffolded tutorials and activities, and it seems intuitive that proper training would result in higher quality questions. The extent to which this is true would be an interesting avenue for future work.

The tutorials on MCQ criticism were exceptionally easy to run and our students provided positive feedback on these experiences. The tutorials on Bloom's Taxonomy were also valuable in getting students to recognise that MCQs can test higher-level thinking rather than just recall, and the classes were highly effective at giving them the frames of reference and vocabulary for criticising peers' questions.

The workload associated with the training, as described in the methods and Figure 1, was not onerous. For example, selecting the set of seven questions for students to discuss in small groups during the initial tutorial took less than an hour. The tutorial itself lasted 1 h, with about 20 min of this being an instructor-led introduction and the remainder involving student discussion of the questions. Beyond this, input from academics is eminently sustainable: the identification and allocation of learning outcomes at suitable granularity, to ensure broad syllabus coverage and to make question selection (Cycle 1) or authoring (Cycle 2) straightforward for students. Using the in-built PeerWise reports and custom dashboards, assessment of these tasks was rapid and feedback communication was easy, encouraging students to give timely, meaningful and actionable feedback to their peers.

Conclusion

We have demonstrated a practical approach for training students in MCQ design and deploying student-authored questions onto a high-stakes summative assessment. With the selection of suitable candidate items resulting from rigorous analysis of performance metrics, the questions provide a defensible assessment of student performance within the unit of study. We found that the selected student-authored MCQs performed as effectively as, and sometimes better than, the academic-authored and textbook-derived MCQs. Leveraging the effort of the student cohort in this fashion represents an opportunity for academics to sustainably build large banks of high-performing, syllabus-aligned MCQs with only a modest impact on workload.

References

Aflalo, E. (2018) 'Students generating questions as a way of learning', *Active Learning in Higher Education*, pp. 1–13. doi: 10.1177/1469787418769120

Amini, N., *et al.*, (2020) 'Inclusion of MCQs written by radiology residents in their annual evaluation: innovative method to enhance resident's empowerment?', *Insights into Imaging*, vol. 11, no. 1, pp. 1–8. doi: 10.1186/s13244-019-0809-4

Bates, S. P., *et al.*, (2014) 'Assessing the quality of a student-generated question repository', *Physical Review Special Topics-Physics Education Research*, vol. 10, no. 2. pp. 1–11. doi: 10.1103/PhysRevSTPER.10.020105

Biggs, J. B. & Tang, C. (2011) *Teaching for Quality Learning at University*, McGraw-Hill Education, Maidenhead.

Bottomley, S. & Denny, P. (2011) 'A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions', *Biochemistry and Molecular Biology Education*, vol. 39, no. 5, pp. 352–361. doi: 10.1002/bmb.20526

Denny, P., et al., (2008) 'PeerWise: students sharing their multiple choice questions', *Proceedings of the Fourth international Workshop on Computing Education Research*, Association for Computing Machinery, New York, NY, pp. 51–58. doi: 10.1145/1404520.1404526

Denny, P., Luxton-Reilly, A. & Hamer, J. (2008) 'Student use of the peerwise system', *SIGCSE Bull.,* vol. 40, no. 3, pp. 73–77. doi: 10.1145/1597849.1384293

Denny, P., *et al.*, (2017) 'Examining a student-generated question activity using random topic assignment', *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, Association for Computing Machinery, New York, NY, pp. 146–151. doi: 10.1145/3059009.3059033

Doyle, E. & Buckley, P. (2020) 'The impact of co-creation: an analysis of the effectiveness of student authored multiple choice questions on achievement of learning outcomes', *Interactive Learning Environments*, pp. 1–10. doi: 10.1080/10494820.2020.1777166

Duret, D., *et al.*, (2018) 'Collaborative learning with peerwise', *Research in Learning Technology,* vol. 26, no. 0, pp. 1–13. doi: 10.25304/rlt.v26.1979

Galloway, K. W. & Burns, S. (2015) 'Doing it for themselves: students creating a high quality peer-learning environment', *Chemistry Education Research and Practice*, vol. 16, no. 1, pp. 82–92. doi: 10.1039/c4rp00209a

Gooi, A. C. C. & Sommerfeld, C. S. (2015) 'Medical school 2.0: how we developed a student-generated question bank using small group learning', *Medical Teacher*, vol. 37, no. 10, pp. 892–896. doi: 10.3109/0142159x.2014.970624

Hancock, D., *et al.*, (2018) 'Improving large class performance and engagement through student-generated question banks', *Biochemistry and Molecular Biology Education*, vol. 46, no. 4, pp. 306–317. doi: 10.1002/bmb.21119

Hardy, J., *et al.*, (2014) 'Student-generated content: enhancing learning through sharing multiple-choice questions', *International Journal of Science Education*, vol. 36, no. 13, pp. 2180–2194. doi: 10.1080/09500693.2014.916831

Harper, R. (2003) 'Multiple-choice questions – a reprieve', *Bioscience Education*, vol. 2, no. 1, pp. 1–6. doi: 10.3108/beej.2003.02000007

Harris, B. H. L., *et al.*, (2015) 'A novel student-led approach to multiple-choice question generation and online database creation, with targeted clinician input', *Teaching and Learning in Medicine*, vol. 27, no. 2, pp. 182–188. doi: 10.1080/10401334.2015.1011651

Hudson, S. L., Jarstfer, M. B. & Persky, A. M. (2018) 'Student learning with generated and answered peer-written questions', *American Journal of Pharmaceutical Education*, vol. 82, no. 2, pp. 96–99. doi: 10.5688/ajpe6315

Jobs, A., *et al.*, (2013) 'Question-writing as a learning tool for students – outcomes from curricular exams', *BMC Medical Education*, vol. 13, no. 1, pp. 89. doi: 10.1186/1472-6920-13-89

Kay, A. E., Hardy, J. & Galloway, R. K. (2018) 'Learning from peer feedback on student-generated multiple choice questions: views of introductory physics students', *Physical Review Physics Education Research*, vol. 14, no. 1, pp. 1–17. doi: 10.1103/PhysRevPhysEducRes.14.010119

Kay, A. E., Hardy, J. & Galloway, R. K. (2020) 'Student use of peerwise: a multi-institutional, multidisciplinary evaluation', *British Journal of Educational Technology,* vol. 51, no. 1, pp. 23–35. doi: 10.1111/bjet.12754

Kelley, M. R., *et al.*, (2019) 'Generation and retrieval practice effects in the classroom using peerwise', *Teaching of Psychology*, vol. 46, no. 2, pp. 121–126. doi: 10.1177/0098628319834174

Masters, J. C., *et al.*, (2001) 'Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education', *Journal of Nursing Education*, vol. 40, no. 1, pp. 25–32. doi: 10.3928/0148-4834-20010101-07

Mccoubrie, P. (2004) 'Improving the fairness of multiple-choice questions: a literature review', *Medical Teacher*, vol. 26, no. 8, pp. 709–712. doi: 10.1080/01421590400013495

Mcleod, P. J. & Snell, L. (1996) 'Student-generated MCQs', *Medical Teacher*, vol. 18, no. 1, pp. 23–25. doi: 10.3109/01421599609040257

Momsen, J. L., *et al.*, (2010) 'Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills', *CBE-Life Sciences Education*, vol. 9, no. 4, pp. 435–440. doi: 10.1187/cbe.10-01-0001

Panczyk, M., *et al.*, (2018) 'Does repetition of the same test questions in consecutive years affect their psychometric indicators? – five-year analysis of in-house exams at Medical University of Warsaw', *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 14, pp. 3301–3309. doi: 10.29333/ejmste/91681

Papinczak, T., *et al.*, (2012) 'Using student-generated questions for student-centred assessment', *Assessment & Evaluation in Higher Education*, vol. 37, no. 4, pp. 439–452. doi: 10.1080/02602938.2010.538666

Purchase, H., *et al.*, (2010) 'The quality of a peerwise MCQ repository', *Proceedings of the Twelfth Australasian Conference on Computing Education – Volume 103*, Australian Computer Society, Inc., Brisbane.

Schullo-Feulner, A., *et al.*, (2014) 'Student-generated, faculty-vetted multiple-choice questions: value, participant satisfaction, and workload', *Currents in Pharmacy Teaching and Learning*, vol. 6, no. 1, pp. 15–21. doi: 10.1016/j.cptl.2013.09.019

Snow, S., *et al.*, (2019) 'A discursive question: supporting student-authored multiple choice questions through peer-learning software in non-STEMM disciplines', *British Journal of Educational Technology*, vol. 50, no. 4, pp. 1815–1830. doi: 10.1111/bjet.12686

Walsh, J. L., *et al.*, (2018) 'Formative student-authored question bank: perceptions, question quality and association with summative performance', *Postgraduate Medical Journal*, vol. 94, no. 1108, pp. 97–103. doi: 10.1136/postgradmedj-2017-135018

Xie, B., *et al.*, (2019) 'An item response theory evaluation of a language-independent CS1 knowledge assessment', *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, Association for Computing Machinery, Minneapolis, MN. doi: 10.1145/3287324.3287370