Positive Impact of Multiple-Choice Question Authoring and Regular Quiz Participation on Student Learning

C. Daniel Riggs,[†]* Sohee Kang,[‡] and Olivia Rennie[†]

[†]Department of Biological Sciences and [‡]Department of Computer and Mathematical Sciences, Centre for Teaching and Learning, University of Toronto, Scarborough, Toronto, Ontario M1C1A4, Canada

ABSTRACT

We previously developed an online multiple-choice question authoring, learning, and self-assessment tool that we termed Quizzical. Here we report statistical analyses over two consecutive years of Quizzical use in a large sophomore-level introductory molecular biology course. Students were required to author two questions during the term and were also afforded opportunities to earn marks for quiz participation. We found that students whose final grade was "A," "B," or "C" exhibited similar patterns of Quizzical engagement. The degree to which students participated was positively associated with performance on formal exams, even if prior academic performance was considered as a covariable. During both terms investigated, students whose Quizzical engagement increased from one exam to the next earned statistically significant higher scores on the subsequent exam, and students who attempted Quizzical questions from earlier in the term scored higher, on average, on the cumulative portion of the final exam. We conclude that the structure and value of the assignment, and the utility of Quizzical as a discipline-independent active-learning and self-assessment tool, enabled students to better master course topics.

INTRODUCTION

University student enrollment has increased dramatically in the past 50 years, and economic factors have conspired to coincidentally increase class size (Douglass and Bleemer, 2018; Roser and Ortiz-Ospina, 2019). In most cases, there has not been an incremental increase in funding, which has led to the widespread use of multiple-choice question (MCQ) examinations. These are particularly prevalent in large introductory classes, as they are economically prudent, and the widespread use of rapid and accurate machine scoring negates grading fatigue/bias that may otherwise be problematic. Moreover, the recent introduction of item analysis has provided instructors with metrics of question quality and the value of distractors, identifying flawed items and guiding decisions on question retention or revision. Thus, while MCQs are not useful for testing higher-order cognitive processes as defined in Bloom's taxonomy of educational objectives (e.g., judgment/creativity; Anderson *et al.*, 2001), they are generally an economically feasible and equitable means of assessment and can be used to measure some of the higher-level objectives of Bloom's taxonomy. Despite the widespread use of MCQ examinations, students report dissatisfaction with this format (Roberts, 1993; Mingo *et al.*, 2018). Anecdotal evidence suggests that many students view some questions as being ambiguous and/or "tricky" (overly specific). Such perceptions may have a foundation that is the fault of the instructor, who may not abide by best practices in developing questions, even though many university faculty development offices offer workshops on the topic and there are many reviews that describe the important features of robust questions (Haladyna *et al.*, 2002; DiBattista and Kurzawa, 2011; Towns, 2014; Moreno *et al.*, 2015; Butler, 2018). The notion of "trickiness" may in part be due to lack of mastery of discipline-specific terminology by students and

Peggy Brickman, *Monitoring Editor*

Submitted Sep 30, 2019; Revised Mar 6, 2020; Accepted Mar 12, 2020

CBE Life Sci Educ June 1, 2020 19:ar16

DOI:10.1187/cbe.19-09-0189

*Address correspondence to: C. Daniel Riggs (riggs@utsc.utoronto.ca).

their resultant inability to distinguish between related answer choices (the distractors). Finally, a negative perception of MCQ examinations may also be due to disconnects that students face between the typical free-flowing dialogue in the classroom versus the terse framing that characterizes most MCQs on exam day. As such, for courses that employ MCQs as the sole or primary means of assessment, it is important for instructors to provide relevant sample questions and for students to prepare in ways that mimic the MCQ format. Indeed, studies have demonstrated that student-authored questions can improve exam performance and provide opportunities for transfer-appropriate processing that improves memory encoding and retrieval (Morris *et al.*, 1977; Lockhart, 2002; Bugg and McDaniel, 2012; McCurdy *et al.*, 2017; Collins *et al.*, 2018; Bae *et al.*, 2019).

It is well recognized that active-learning and experiential-learning strategies enhance comprehension and promote deeper learning across many disciplines (Prince, 2004; Ruiz-Primo *et al.*, 2011; Freeman *et al.*, 2014; Holley, 2017; Ott *et al.*, 2018; Schroeder *et al.*, 2018). A meta-analysis of 225 studies on science, technology, engineering, and mathematics (STEM) courses demonstrated that active-learning exercises increased exam performance by half a letter grade and were associated with increased retention of students (Freeman *et al.*, 2014). Active learning can take many forms but ultimately involves students participating in activities that promote engagement and use of course materials. Ideally, this involves dialogue between pairs or groups of students and is designed to assist learners in understanding information and critically evaluating it by posing and answering questions. In this regard, there is a large body of evidence that links question/problem authoring to enhanced exam/course performance (Draper, 2009; Hardy *et al.*, 2014; Schroeder *et al.*, 2018), and a variety of Web-based applications can be employed to facilitate engagement. Some examples are the class response system CLICKERS (Martyn, 2007), the MCQ authoring system PeerWise (McQueen *et al.*, 2014; Hardy *et al.*, 2014), and concept-mapping software (Weinertha *et al.*, 2014).

We have developed an online MCQ authoring, testing, and learning tool called Quizzical (Riggs *et al.*, 2014). This turnkey application is Learning Tools Interoperability (LTI) compliant and, upon setup, the instructor can select from many options to tailor the authoring, grading, and quiz functions. We deployed the software in a sophomore-level molecular biology class, requiring students to author two questions during the term, and students were awarded participation marks for timely engagement in answering questions related to each class. Our premise was that requiring deeper engagement with the topics and providing unlimited access to the generated quiz bank would afford students opportunities for deeper learning and exam preparation. Our statistical analyses reveal positive correlations between Quizzical use and exam/course performance.

METHODS

Quizzical Overview

Given that formal testing in large courses is a daunting task, with equity in grading, timely feedback, and limited resources being considerations/constraints, we have long used MCQs as the format for all exams. We wanted students to have access to relevant practice questions without compromising the instructor's test bank, and an online MCQ authoring and quiz tool called Quizzical was developed to facilitate this goal and to promote learning (Riggs *et al.*, 2014). Quizzical 3.0 is an LTI-compliant application that should work seamlessly within the secure environment of most commonly used learning management systems. As such, students log in to their accounts and connect directly to Quizzical through the relevant course page. Quizzical has two major features: student authoring of MCQs and the use of these questions by students as learning tools and to get relevant practice quizzes for formal examinations. These aspects are more fully described in the Student Stakeholder section below.

The Instructor Stakeholder and Quizzical Options

Deploying Quizzical is straightforward, as it was designed for instructors with little or no experience in using educational software. The initial setup is structured to guide the new user through a series of steps to schedule student and teaching assistant (TA) assignments, to upload images from other sources (e.g., a textbook image library) if desired, and to choose options to customize a course. Once the course has been set up, there is little or no intervention needed. We have used Quizzical for several years and make use of both the authoring and quiz functions for our students. End-of-term course evaluations and written comments suggest that the application is highly valued by students (Riggs *et al.* 2014). Based on our experience and the options that we have found useful, a flowchart of the authoring and quiz timelines is presented in Figure 1. It should be noted that instructors can encourage participation and course engagement by requiring students to attempt quizzes during the term. In our case, we allowed students to earn as much as 6% of their final course grades by meeting the following criteria: For each lecture, attempt at least 10 questions within 14 days of the lecture, scoring at least 60%.

The instructor's dashboard (Supplemental Figure S1) contains 10 tabs at the top to allow the instructor to conduct the setup process or change parameters of the course (e.g., reschedule a student assignment), monitor student and TA activities, view and obtain statistical information about questions, and manage the grade book. Filterable graphs permit analysis of questions, student performance, and TA progress. A video that illustrates all features of Quizzical can be found in the supplemental files (Supplemental Video S1) or via the University of Toronto link: https://play.library.utoronto.ca/kIYIj0Ni3HVl.

The Student Stakeholder: Dashboard and Overview of Authoring/Quiz Functions

The student dashboard is shown in Supplemental Figure S2. Alerts for upcoming authoring and quiz participation obligations are prominently shown. The Questions box shows the status of the student's questions, including lecture number, due date, status, grader, and earned grade. Action buttons permit the student to view completed work or take the student to another page containing a form filler for authoring a question. Quiz attempts are shown graphically, capturing the number of attempts and the percent success on both a per-lecture basis and chronologically. Students can chart their progress and evaluate their understanding of the topics for each lecture. At the top are two tabs: Writing Effective Questions, which takes the student to a summary of the best practices for authoring questions and provides a synopsis of Bloom's taxonomy of educational objectives along with some appropriate verbs to use for each category.

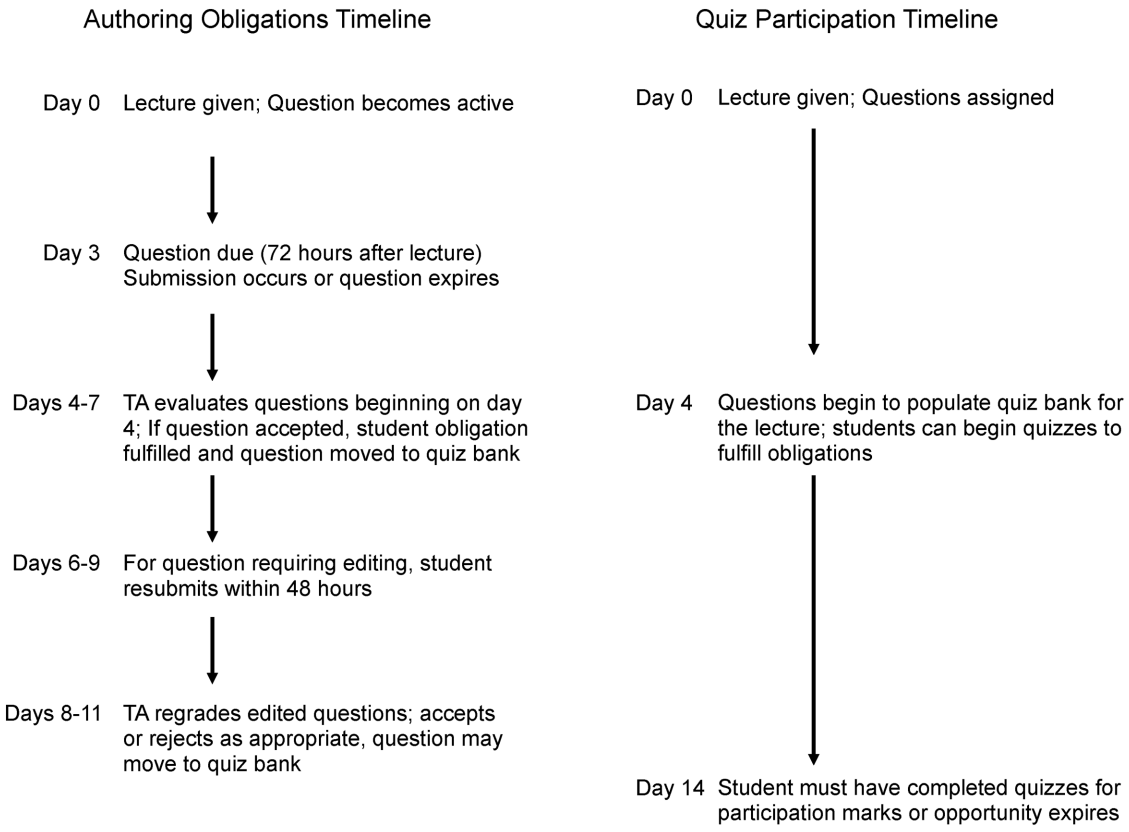

FIGURE 1. Flowchart for Quizzical authoring and quiz participation. During course setup, the instructor determines the interval of time between a formal lecture and the due date for student-authored questions to be submitted, whether editing and resubmission of a flawed question is permitted, and the duration of time when quizzes can be undertaken to earn participation marks (if earning participation marks is allowed). In this example flowchart, the lecture occurs on day 0, and the author has 3 days to submit. TAs begin to grade questions after day 4 and are encouraged to complete their work within a few days. If editing is permitted, a flawed question is returned to the student author with guidance for corrections and can be resubmitted within 2 days of grading (or it expires). With regard to quiz participation, note that, given the deadlines for authoring, quiz questions do not begin to populate the quiz bank until day 3–4. If the quiz participation deadline is 14 days, this affords students ~7–10 days to take quizzes to earn participation marks.

The other tab, Take a Quiz, links to another page that allows the student to choose a lecture for taking a quiz and presents a tabular summary of performance for all lecture quiz pools. Students can log in at any time to take quizzes, and once the question pool for a particular lecture is exhausted, the questions are then randomized and provided for subsequent reattempts.

When an authoring assignment becomes active (on the date of a lecture), the student clicks on Compose, and a form filler is presented (Supplemental Figure S3). Several help icons have drop-down descriptions to assist the student. The student must categorize his or her question as being either a Recall knowledge or an Application question, emphasizing different levels of Bloom's taxonomy. There are boxes in which the question, the answer, and the distractors are typed. Depending on the options selected by the instructor, the student may be required to provide other information that is not presented to quiz takers until they have attempted the question. These resources are designed to help quiz takers better understand the question and include providing a reference (e.g., "page 343," "Figure 10-17"), selecting an image to associate with the question (e.g., from a set of figures uploaded by the instructor from the textbook), and providing justifications for both the correct answer and the distractors.

When taking a quiz, the student is presented with one question at a time and must select an answer by clicking on the appropriate radio button (Supplemental Figure S4). Once an answer is submitted, Quizzical indicates whether the student was correct or incorrect and provides the output page (Supplemental Figure S5) containing the elements alluded to earlier. The inclusion of the justifications and an image affords students a learning opportunity, as this strategy provides them with rationales for why a particular answer is right or wrong. A more thorough description of the student dashboard, resources, authoring, and quizzing is provided in the accompanying Supplemental Video S1 or via the University of Toronto link: https://play.library.utoronto.ca/6XzobNLPS8jx.

Participants and Context

We have employed Quizzical for several years at the University of Toronto in a sophomore-level introductory molecular biology class of approximately 500 students. In this paper, we present data for the 2017 and 2018 terms, when the constants included the same instructor, textbook, and topics. There were no curricular changes that influenced the participants. In both of these years, each student was required to author two questions

during the term, and submission dates were randomly determined by the Quizzical algorithm. Each student-authored question was worth 4% of the student's final grade. In addition, students were encouraged to participate in taking quizzes by leveraging the test bank of approved questions. As many studies have demonstrated value to reviewing topics shortly after their introduction, and there is a negative relationship between procrastination and performance (reviewed by Kim and Seo, 2015), we employed a formula for participation marks whereby full credit would be given for attempting at least 10 questions and scoring at least 60% within 14 days of the class date. In 2017, students could earn 8% of their final grades in this way, whereas in 2018, the value was reduced to 6%. In both years, the class consisted of 24 formal lectures, and three examinations were held. Term test 1 (TT1) assessed mastery of lectures 1–8, term test 2 (TT2) focused on lectures 9–16, and term test 3 (TT3, the noncumulative portion of the final exam) focused on lectures 17–24. There was a cumulative component to the final exam that assessed mastery of topics in lectures 1–16.

Intervention, Data Collection, and Data Analyses

The class began with a tutorial, largely centered on familiarizing students with the Quizzical interface and dashboard, the authoring/participation components, and reviewing best practices for writing effective MCQs (see also Supplemental File S1 for a description of the tutorial components). A second intervention involved interactions between the student author and the grader (either the instructor or a TA, if specified during setup). Upon submission of a question by a student author, Quizzical routes the question to a TA for evaluation (see also Supplemental File S1 for a summary of instructions given to TAs). If the question satisfies the instructor's rubric and scoring threshold, it may be directly approved by the TA, marks are recorded, and the question is deposited into the quiz pool for the relevant lecture. If the TA finds issues with the question, he or she may give the student author feedback/guidance to allow the student to appropriately edit the question for resubmission. This shows up as an alert on the student dashboard. In terms of a rubric, for a total of 4 marks, we allocated 1 mark for on-time submission, 0.5 marks for citing an appropriate reference/textbook image, and 0.5 marks for proper categorization of the question as recall or application and proper verb selection. The final 2 marks were at the discretion of the TA for the adherence to best practices for the design of the question and for the strength of the justifications of the answer/distractors.

On the initial log-in, students are asked for their permission to have their questions used for quiz purposes, as potential test questions, and for other future pedagogical analyses for the class. The information security, potential conflicts, potential risks and benefits, confidentiality and privacy, and debriefing/dissemination aspects were reviewed and approved by the University of Toronto research ethics board (Protocol 35875). Quizzical monitors quiz-bank usage, recording a time stamp, the number of questions attempted, and the number of questions answered correctly. These three bits of data are used to calculate the participation mark for each lecture, should the instructor elect to use this feature. Quizzical also identifies activity for each question, such that the global number of attempts and percent correct are recorded and used to calculate a point biserial score (Pearson product-moment correlation coefficient) for the question to gauge its value. Finally, after attempting a question and then being presented with the answer/justifications, students are asked to rate the question on a Likert scale from 1 (poor) to 5 (excellent) stars. This value is displayed to the instructor and may be considered a student value-perception rating. Question metrics can be seen under the Report tab on the instructor's dashboard (Supplemental Figure S1 and described in the Instructor's Guide video at https://play.library.utoronto.ca/kIYIj0Ni3HVl). We used the point biserial score and the student rating score to select student questions for formal assessments. On average, about 15% of the exam questions were student-authored or modifications of them. Item analysis from each formal exam indicated that each Quizzical question earned biserial scores that were both consistent with the exam and qualified as being highly discriminatory (biserial values 0.25–0.38).

The statistical software, R (www.r-project.org) was used for all statistical analyses and generation of figures. Data tables for all analyses are available as Supplemental Material.

Quizzical Availability

Instructors who wish to employ Quizzical at their institutions should contact the corresponding author for details.

RESULTS

To begin to dissect whether Quizzical use is beneficial to students, we assembled the usage metrics for two consecutive terms. Table 1 shows that, for both years, the class began with approximately 500 students and suffered a 20–30% attrition rate. For both years, there were significantly more female students. Students were required to author two questions during the term, and in total, more than 740 questions were authored in both terms. Given that there were 24 classes per term, this resulted in approximately 30–35 questions being authored on the array of topics covered in each class, and on average, students attempted over 400 questions each during the term.

We first examined Quizzical daily quiz-bank usage during these two terms. Because both patterns were very similar, only the winter 2017 graphs are displayed. Figure 2A shows the sum of attempts throughout the term. There are three spikes of activity near the term tests and the final exam dates. These peaks suggest students were active users for self-assessment just before formal exams. In addition, the graph reveals a repeating double peak, which we interpret as students fulfilling their obligations for participation marks by completing 10 questions for each lecture within a 14-day period (see *Methods* section for details). This is more obvious when the attempts are restricted to valid participation attempts (i.e., the student met the criteria for earning participation marks; Figure 2B). We also examined the median number of attempts by each date per student. The minimum and maximum values of median attempts per day

TABLE 1. Quizzical metrics for 2017 and 2018

Year	No. of students (F/M)[a]	No. of questions[b]	No. of attempts[c]
2017	380 (233/147)	789	188,646
2018	355 (235/120)	740	148,471

[a]Total number of students and female/male ratio at the end of the course.
[b]Total number of questions generated.
[c]Total number of questions attempted.

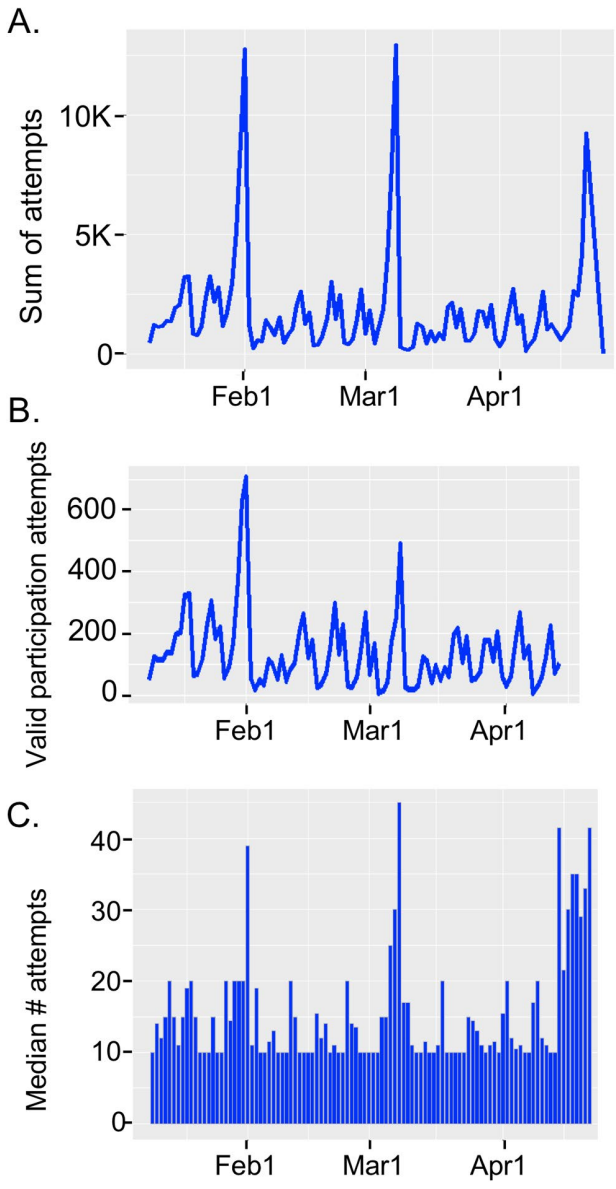

FIGURE 2. Daily Quizzical use patterns in Winter 2017. (A) The daily sum of the total number of attempts was plotted, revealing three distinct peaks of activity that coincided with the two term tests and the final exam (February 1, March 8, and April 21 dates). (B) The daily sum of attempts, corrected for valid participation marks only, more clearly reveals that students generally met the criteria (attempt 10 questions per lecture within 14 days of the date, scoring at least 60%). The repeating double peaks correspond to the two lectures held each week. (C) Median number of attempts per student throughout the term.

were 10 and 45, respectively, with a mean of 15.29 and an SD of 7.87 median number of attempts per day (Figure 2C).

We hypothesized that student engagement, as measured by the median number of questions attempted versus final course grade, would show that high achievers would attempt more questions and show more consistent participation than low achievers. Figure 3 shows graphs for students earning each of the letter grades "A" to "D" and "F." Surprisingly, the "F" students attempted the most questions, followed by the "A" students, with "B," "C," and "D" students exhibiting similar but lower engagement levels. In terms of usage trends, the "A" to "C" students were very similar in structured engagement, seemingly fulfilling the requirements for participation marks and also using Quizzical more intensely in the days preceding an exam (test practice). The "A" students exhibited higher levels of test practice versus the other groups, with the exception of the "F" students, who were less consistent in meeting the obligations for participation marks throughout the term, focusing instead on intense use just before an exam. Thus, the "F" students were outliers from a general trend of regular engagement and intense test practice. Interestingly, the "A" to "C" students' test practice generally increased for the second and third exams, suggesting that students saw value in this enterprise, while the "D" and "F" students displayed more random and/or erratic patterns of usage. The extreme level of engagement by "F" students after the April final exam we believe to be related to them trying to earn participation marks, as the exam was scheduled less than a week from the final class and students still had another week to earn participation marks for the final two lectures.

Impact of Quizzical Engagement on Final Course Grade

We explored whether Quizzical engagement had a positive impact on students' course performance in two ways: one wherein 2 years of data (2017 and 2018 classes) were combined and one wherein data from each year were analyzed separately. We used the participation score, wherein full marks were awarded for meeting the criteria of attempting at least 10 questions per lecture within 14 days of the lecture and scoring at least 60%. A pro rata formula awards marks for participation that falls short of this threshold. We categorized Quizzical engagement in three ways: Low (students in less than the 25th percentile of engagement), Mid (between the 25th to 75th percentile), and High (higher than the 75th percentile). For this analysis, a student's final grade was computed as the average of the first and second term test scores, plus the score on the noncumulative portion of the final exam (i.e., material not covered on either of the previous term tests). Figure 4 shows box plots for both years analyzed, and Tukey's multiple comparisons revealed highly significant mean differences for all pairwise comparisons (students in all three levels of Quizzical engagement differ significantly from peers in other levels of engagement; see Supplemental Table S1-1 and S1-2). We combined the data for both years and ran an analysis of covariance (ANCOVA) test for the final grade with the Quizzical engagement categories, controlling for each year and prior academic performance (incoming grade point average [GPA] of students). We found that there was a highly significant course performance difference between the three categories ($p = 3.037 \times 10^{-14}$; see Supplemental Table S1-3), one that is not based on previous academic ability, and that there was no significant difference between the 2 years.

We hypothesized that both incentivizing participation and requiring students to submit questions would correlate with higher academic performance. We combined data for the 2 years and used incoming GPA as a covariate to assess the value of authorship only. ANCOVA revealed that there are highly significant course performance differences by the different levels of authorship after we control for previous academic performance

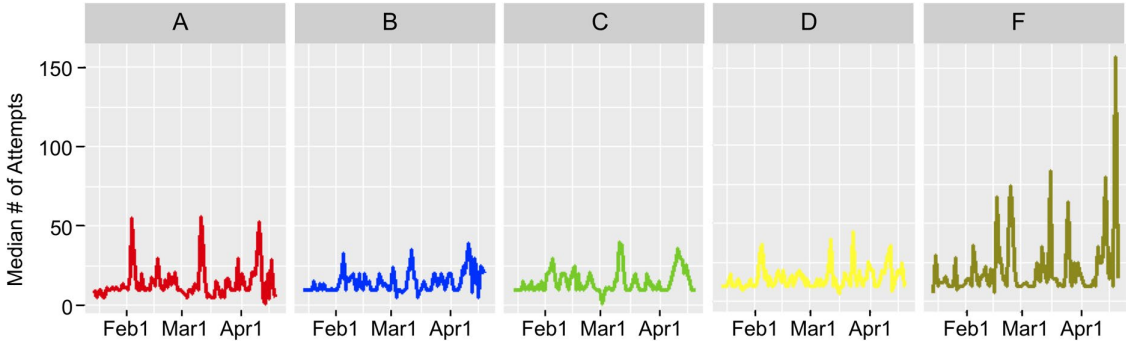

FIGURE 3. Median number of attempts by students in each grade category. Final course grade was used to sort students into groups of "A," "B," "C," "D," and "F" categories, and daily engagement was then determined by plotting the median number of questions attempted. Note that, in this term (2018), the formal exam dates were February 5, March 12, and April 11, which coincide with peaks of elevated use. A similar pattern was observed for the 2017 class (unpublished data).

($p = 2 \times 10^{-16}$, $F = 98.6$; see Supplemental Table S1). Similarly, when scored in combination with authorship, quiz participation was highly correlated with test performance ($p < 2.2 \times 10^{-16}$, F value = 187.6; see Supplemental Table S1-3).

Changes in Engagement during the Term Are Correlated with Changes in Test Performance

We next asked whether a change in the level of Quizzical engagement between term tests impacted students' performance on the subsequent term test. That is, if a student increased his or her Quizzical use between TT1 and TT2 (relative to use before TT1), would his or her performance improve on TT2 (vs. TT1)? For each examination (TT1 and TT2 plus the noncumulative final exam material), we measured whether changes in Quizzical engagement impacted exam performance in each of the 2 years. Quizzical engagement was defined as the quiz participation scores that students earned. Based on performance improvement or decline between examinations, we defined two groups: group 0, for whom Quizzical engagement decreased (e.g., less participation in the time period between TT1 and TT2 or between TT2 and final exam); and group 1, for whom Quizzical engagement increased in these time intervals. Figure 5 shows that, for both years, group 1 students' increased Quizzical participation resulted in a higher subsequent term test performance. The t tests among the two groups for each test and for both years were all statistically significant (the largest p value was 0.04, and the smallest p value was 0.002; see Supplemental Table S2-1 for the summary table). This result was consistent when we performed ANCOVA, using combined data from 2 years on TT2 and TT3, and after controlling for prior academic knowledge (previous cumulative GPA), there were still highly significant mean differences between the two Quizzical engagement groups on both term tests (see Supplemental Table S2-2).

Review of Early Course Topics Promotes Higher Scoring on the Cumulative Final Exam

Given the strong correlations between Quizzical use and subsequent test scoring, it might be expected that review of questions on previous term test material (TT1 and/or TT2) would be positively correlated with higher performance on TT1/TT2 questions on the final cumulative exam. For these analyses, we recorded TT1 (lectures 1–8 during weeks 1–4) and TT2 (lectures 9–16 during weeks 5–8) question attempts in the 3-week period (weeks 10–12), preceding the final exam and compared this activity with scores on TT1 and TT2 questions on the final exam. Because fewer students attempted this type of review, we combined data for both years and used the students' term test average as a covariate. Regression analyses revealed that students who undertook TT1 question review scored on average 2.8% higher than those who did not review, and

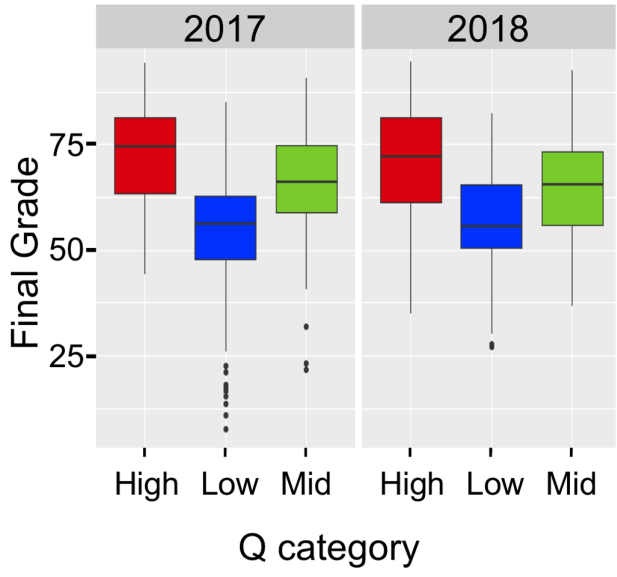

FIGURE 4. Box plots of Quizzical engagement categories vs. formal exam averages. The participation marks were calculated as described in the Figure 2 legend and the text, and three performance categories were used to model engagement. "High" represents the top 25th percentile, "Mid" represents the 25th–75th percentile, and "Low" represents the lower 25th percentile categories of student engagement, plotted against the average of the formal exams. Tukey's multiple-comparison test was employed, and adjusted p values were calculated. All pairwise comparisons were found to be statistically significant with p values $< 3 \times 10^{-7}$. Including incoming GPA as a covariate did not influence the degree of significance (see Supplemental Table S1 for ANCOVA table).

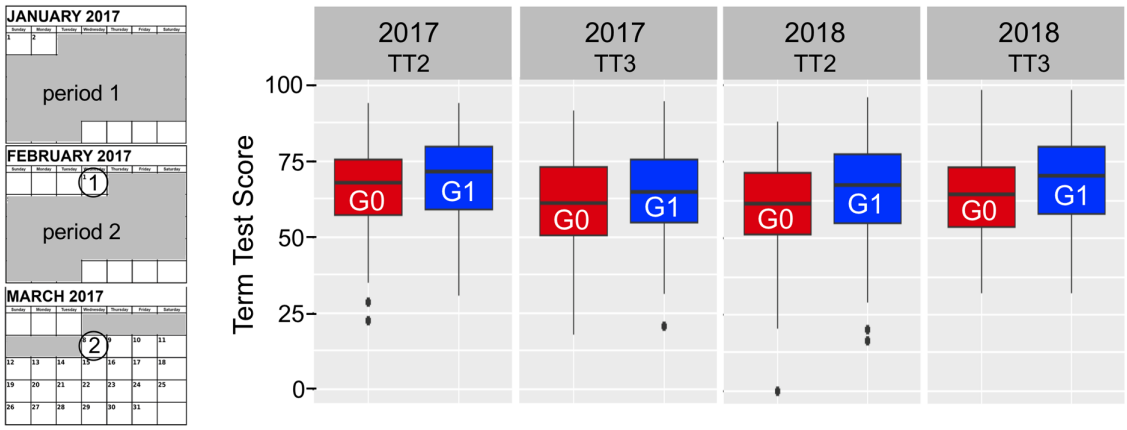

FIGURE 5. Increased engagement is correlated with improved exam performance. Two time periods of Quizzical participation were monitored, and their relationship to two term test dates (circled) are shown on the left (2017 data only, as an example). The second panel shows box-plot data related to term test 2 scores in 2017. Students who participated less often in period 2 compared with period 1 are group 0 (G0), whereas students who participated more in period 2 vs. period 1 are group 1 (G1). The other box plots show data for comparable periods for the final exam (2017 TT3) and comparable data from 2018 (plots 3 and 4). Paired *t* tests revealed statistical significance for all pairwise comparisons of G0 and G1 data (*p* values are 0.03929, 0.0212, 0.0038, and 0.002, respectively).

ANCOVA revealed robust statistical significance when the students' formal test achievement was considered ($p = 3.037 \times 10^{-14}$, $F = 59.9$; see Supplemental Table S3). Similarly, students who reviewed TT2 questions scored on average 3.2% higher than those who did not (ANCOVA p value $< 2.2 \times 10^{-16}$, $F = 95$; see Supplemental Table S3).

In summary, we employed several statistical analyses to show the impact of Quizzical engagement on test and course performance. The results presented here provide compelling evidence that Quizzical use strongly supports students' learning in the course, after controlling for prior academic ability.

DISCUSSION

An impediment to potential users incorporating new educational software is the real or perceived steep learning curve in making it operational and in dealing with unknowns. The latest version of Quizzical features a course setup that guides instructors through options for student authoring, assignment number and scheduling, TA tasks, grading, participation marks, uploading of images, and quiz parameters. It may be considered a turnkey application that requires very little management time and few computer skills. There are two options that we feel are very beneficial to students: rewarding quiz participation marks and answer/distractor validation by authors. In the initial Quizzical trials, we did not designate participation in quizzes as a component of the final grade. More recently, we employed a strategy whereby students could earn 6–8% of the course grade by fulfilling the criteria of: within 14 days of each lecture, attempting at least 10 questions and scoring at least 60%. We implemented this strategy because many studies have shown that knowledge is best assimilated if engagement occurs soon after its initial presentation and procrastination leads to negative academic performance (Michinov *et al.*, 2011; Kim and Seo, 2015). Retrieval practice, wherein students are formally assessed, or they self-assess through activities like online quizzes, is critically important to mastering new information and encoding it into long-term memory (Karpicke and Blunt, 2011; Roediger and Butler, 2011; Dunlosky *et al.*, 2013). Students with higher levels of engagement performed better on exams, and this is likely due in part to the spacing of retrieval practice. For example, Dobson and coworkers (2017) reported that students performed much better on exams when employing a distributed study/retrieval strategy for learning anatomy versus other approaches that involved more intensive study and/or deferred retrieval. Moreover, our success with implementing the participation marks may be considered an example of transfer-appropriate processing of the testing effect, whereby student performance is positively affected by prior engagement in activities that model the style and conceptual framework of summative examinations (Shaibah and van der Vleuten, 2013; Nguyen and McDaniel, 2015; Collins *et al.*, 2018; Bae *et al.*, 2019). Anecdotally, on end-of-term student evaluations, a number of students commented that the Quizzical question bank allowed them to learn, review, and self-test in a format that mimicked that of the formal exams. Thus, the participation incentive is likely to be a strong contributor to the positive association we see between engagement and test scoring, and similar correlations have been reported for students using other MCQ software (e.g., PeerWise: Walsh *et al.*, 2018).

A second option that we employed was to require student authors to justify all answers, both the correct answer and the distractors, and to associate textbook images with their questions. For student authors, articulating a rationale for why answers are correct or incorrect promotes deeper engagement and comprehension of the topic (Larsen *et al.*, 2013; Koretsky *et al.*, 2016). Also, when taking a quiz, students see these resources after attempting a question, allowing them to understand the limitations of their knowledge and/or to better contextualize the content to expand their comprehension. This is particularly applicable to visual learners. As such, this immediate feedback affords a learning opportunity and a means to recognize misconceptions (Butler and Roediger, 2008; Hardy *et al.*, 2014; Koretsky *et al.*, 2016; Mullet and Marsh, 2016). We should emphasize that in a large class setting where multiple students are asked to author questions for each lecture, there are often multiple questions about the same topic. While at face value this

may seem repetitive to the quiz taker, a recent study reported that retrieval of information from multiple perspectives enhances long-term information retention (Zheng *et al.*, 2016). Similarly, when pairs of students were asked to collaborate on questions that differed lexically, they tended to engage in more discourse about the questions and performed better on examinations (Jucks and Paus, 2013). Thus, multiple questions on a topic may better prepare those who more actively engage in quizzes.

Interestingly, we found that males performed significantly better than did females in both years of our study (unpublished data). Similar observations for exam and/or course performance have been reported by a number of studies in STEM disciplines (reviewed by Eddy and Brownell, 2016), but meta-analyses have suggested that females outperform males (Voyer and Voyer, 2014; O'Dea *et al.*, 2018). The underlying reasons for these apparent gender biases are not clear, as different studies employ different metrics, assessment formats, and methods of analysis, but there are undoubtedly sociocultural and psychological factors involved. Despite the fact gender bias exists for the course grade in our study, we found no gender differences in Quizzical participation (p value = 0.6689). It is possible that the online nature of Quizzical allows students to participate in a comfortable setting of their choosing, limiting underlying causes such as stereotype threat and test anxiety, but some aspect(s) of formal testing contributes to grade stratification. There is evidence that gender bias may be due to the nature of testing, as it exists for MCQ format exams but is not prevalent in comparable groups where constructed-response questions were the focus of assessments (Stanger-Hall, 2012; Wright *et al.*, 2016). We propose that the relatively low-stakes value of the Quizzical assignment may mitigate negative influences of MCQ testing, promoting greater student acceptance of its value as a learning and review tool.

CONCLUSIONS

Statistical analyses over 2 years of Quizzical use revealed that the options we employed were effective in promoting student learning, and participation was associated with elevated exam scores. Of note, males outperformed females, on average, for both cohorts under study. The gender bias we observed will require further research to evaluate the underlying causes of this outcome. From our results, we conclude Quizzical is a powerful learning tool with the potential for widespread use throughout academia. While further studies are required to establish a more nuanced understanding of the ways in which Quizzical can benefit retention of material, motivation, course enjoyment, and the observed sex bias, this first investigation into its utility is encouraging. Based on these results, it is evident that Quizzical and Quizzical-like learning platforms could become a key component of large courses across academic disciplines.

ACKNOWLEDGMENTS

We thank Zoran Piljevic and Syed Kashif of the University of Toronto Scarborough (UTSC) Information and Instructional Technology Services for programming and computer support services and Adon Irani of the Centre for Teaching and Learning (CTL) for assistance with development and troubleshooting. We acknowledge funding from CTL and the University of Toronto Instructional Technology Innovation Fund to support Quizzical development. We appreciate the constructive criticisms of the anonymous reviewers.

REFERENCES

Anderson, L.W., Krathwohl, D.R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Cognition*, *60*, 206–214.

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology*, *104*, 922–931.

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

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616.

Collins, D. P., Rasco, D., & Benassi, V. A. (2018). Test-enhanced learning: Does deeper processing on quizzes benefit exam performance? *Teaching Psychology*, *45*, 235–238.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), ar4.https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Dobson, J. L., Perez, J., & Linderholm, T. (2017). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education*, *10*, 339–347; https://doi.org/10.1002/ase.1668

Douglass, J. A., & Bleemer, Z. (2018). *Approaching a tipping point? A history and prospectus of funding for the University of California*. Berkeley, CA: Berkeley Center for Studies in Higher Education.

Draper, S. W. (2009). Catalytic assessment: Understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology*, *40*, 285–293. doi: 10.1111/j.1467-8535.2008.00920.x

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58.

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physical Education Research*, *12*, UNSP 020106.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okorafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences USA*, *111*, 8410–8415.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–333.

Hardy, J., Bates, S. P., Casey, M. M., Galloway, K. W., Ross, K., Kay, A. E., ... & McQueen, H. A. (2014). Student-generated content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*, *36*, 2180–2194.

Holley, E. A. (2017). Engaging engineering students in geoscience through case studies and active learning. *Journal of Geoscience Education*, *65*, 240–249.

Jucks, R., & Paus, E. (2013). Different words for the same concept: Learning collaboratively from multiple documents. *Cognition and Instruction*, *31*, 227–254.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

Kim, K. R., & Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality & Individual Differences*, *82*, 26–33.

Koretsky, M. D., Brooks, B. J., & Higgins, A. Z. (2016). Written justifications to multiple-choice concept questions during active learning in class. *International Journal of Science Education*, *38*, 1747–1765, doi: 10.1080/09500693.2016.1214303

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*, 674–682

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*, 397–403. doi: 10.1080/09658210244000225

Martyn, M. (2007). Clickers in the classroom: An active learning approach. *EDUCAUSE Quarterly (EQ)*, *30*(2), 71–74.

McCurdy, M. P., Leach, R. C., & Leshikar, E. D. (2017). The generation effect revisited: Fewer generation constraints enhances item and context memory. *Journal of Memory and Language*, *92*, 202–216.

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

Michinov, S., Bruot, O., Le Bohec, O., Juhel, J., & Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education*, *56*, 243–252.

Mingo, M. A., Chang, H., & Williams, R. L. (2018). Undergraduate students' preferences for constructed versus multiple-choice assessment of learning. *Innovative Higher Education*, *43*, 143–152. doi: 10.1007/s10755-017-9414-y

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple-choice items. *Psicothema*, *27*, 388–394. doi: 10.7334/psicothema2015.110

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. doi: 10.1016/S0022-5371(77)80016-9

Mullet, H. G., & Marsh, E. J. (2016). Correcting false memories: Errors must be noticed and replaced. *Memory & Cognition*, *44*, 403–412.

Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, *42*, 87–92.

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications*, *9*, 3777. doi: 10.1038/s41467-018-06292-0

Ott, L. E., Carpenter, T. S., Hamilton, D. S., & LaCourse, W. R. (2018). Discovery learning: Development of a unique active learning environment for introductory chemistry. *Journal of the Scholarship of Teaching and Learning*, *18*, 161–180. doi: 10.14434/josotl.v18i4.23112

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, *93*, 223–231.

Riggs, C. D., Sutherland, B., Irani, A., & Patterson, J. C. (2014). "Quizzical": A student-authored, on-line, multiple choice question writing and learning assessment tool. Paper presented at: International Conference for Educational Development (Stockholm, Sweden).

Roberts, D. (1993). An empirical study on the nature of trick test questions. *Journal of Educational Measurement*, *30*(4), 331–344.

Roediger, H. L., 3rd, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, *15*, 20–27.

Roser, M., & Ortiz-Ospina, E. (2019). *Global rise of education*. Retrieved August 25, 2019, from https://ourworldindata.org/global-rise-of-education

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & Shepard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, *331*(6022), 1269–1270.

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., & Adesope, O. O. (2018). Studying and constructing concept maps: A meta-analysis. *Educational Psychology Review*, *30*, 431–455.

Shaibah, H. S., & van der Vleuten, C. P. M. (2013). The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy, *Anatomical Sciences Education*, *6*, 149–156.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*, 294–306.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, *91*, 1426–1431. dx.doi.org/10.1021/ed500076x

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*, 1174–1204.

Walsh, J. L., Harris, B. H. L., Denny, P., & Smith, P. (2018). Formative student-authored question bank: Perceptions, question quality an association with summative performance. *Postgraduate Medical Journal*, *94*, 97–103.

Weinertha, K., Koeniga, V., Brunner, M., & Martina, R. (2014). Concept maps: A useful and usable tool for computer-based knowledge assessment? A literature review with a focus on usability. *Computers & Education*, *78*, 201–209

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in an introductory biology course. *CBE—Life Sciences Education*, *15*, 1–16.

Zheng, J., Zhang, W., Li, T., Liu, Z., & Luo, L. (2016). Practicing more retrieval routes leads to greater memory retention. *Acta Psychologica*, *169*, 109–118. https://doi.org/10.1016/j.actpsy.2016.05.014