

RESEARCH

Free and Open Access

# Impact of reusing question banks on test integrity and student learning

Raed Jarrah <sup>1\*</sup> and Ibrahim Itani <sup>2</sup>

\*Correspondence:  
[raed.jarrah@emich.edu](mailto:raed.jarrah@emich.edu)  
School of Engineering,  
Eastern Michigan University,  
900 Oakwood St, Ypsilanti,  
Michigan 48197, USA  
Full list of author information is  
available at the end of the article

## Abstract

Instructors are increasingly utilizing the convenience of computer-based testing to produce randomized versions of examinations to maintain assessment integrity. As preparing a sufficiently large bank of questions for examinations can be very demanding, instructors may be tempted to utilize the same question bank for both practice quizzes and examinations. However, as some of the questions could be exposed during practice, sharing banks would raise concerns about the integrity of the examination. This study developed a formula for calculating the expected integrity of an examination's question bank and then used non-parametric statistical tests to analyze the benefits and risks of sharing question banks between assessments. The study found that sharing banks can still provide learning benefits without significantly impairing the authenticity of an examination. The study also showed that a bank's integrity was not correlated with student performance, question exposure alone did not lead to improved performance, and learning from practice questions improved students' examination performance.

**Keywords:** Computer based assessments, Question banks, Multiple-choice, Academic integrity, Academic misconduct

## Introduction

Teaching and assessment form a crucial feedback loop in higher education. Instructors invest considerable time and effort in crafting assessments that evaluate students' understanding and mastery of course material. While proctoring would reduce academic misconduct among students taking a test, it does not entirely prevent them from memorizing and leaking the questions to students who will take the exam in the future. To safeguard the integrity of examinations, educators often resort to offering various versions of tests to students.



© The Author(s). 2025 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

In many cases, instructors may consider offering different versions of the same exam to thwart academic misconduct among students. Computer Based Testing (CBT) allows for the convenience of generating multiple exam versions from a larger set of prepared questions, often referred to as question banks.

Modern learning management systems also allow instructors to tailor different formats of questions, such as multiple-choice, sorting, matching, calculation, and essays. Apart from essay-type questions, CBT provides the convenience of automated grading. Instructors can also anticipate incorrect responses and provide automated feedback when setting up tests, such as explaining why a selected choice on a question is incorrect.

Despite the convenience of CBT, crafting test questions remains time-consuming, so some instructors may reuse previous questions in their assessments. Moreover, a meticulously designed assessment would have gone through several iterations based on previous student responses, with each revision refining clarity, balancing difficulty, and reducing bias. Nonetheless, instructors must maintain a sufficiently large bank of questions to minimize the chance of a compromised assessment.

To aid students in learning, instructors may assign regular formative assessments in the form of homework exercises or quizzes. Short, in-class quizzes can have the advantage of proctoring, and maintaining the integrity of the assessment. However, such quizzes would also require the formulation of multiple questions to maintain integrity, leading to an increased burden on the instructor in addition to preparing the summative examination. Accordingly, an instructor may be tempted to save some effort by using some of the prepared summative examination questions in these formative assessments. The purpose of this study was to analyze the viability of such a compromise.

This study aimed to identify a mathematical formula to determine the required size of a question bank to maintain a sufficient level of integrity, and then to validate it by analyzing examination data from a university course. The study also aimed to determine if reusing questions in practice quizzes can benefit students without compromising summative examinations. The study's main research questions can be summarized as:

- 1) How many questions should a bank have to achieve a desired integrity?
- 2) Does exposing questions ahead of an assessment reduce its authenticity?
- 3) Does allowing repeated exposure to questions benefit student learning?

## **Literature review**

### **Computer-based testing**

Computer-based examinations have become more accepted by students (Boevé et al., 2015; Ilgaz & Afacan Adanır, 2020). A study found that Australian university students from computer-based disciplines (such as computer science) tended to favor computer-based

exams, while students from disciplines that are more reliant on multiple calculations and diagrams tended to be much less enthusiastic about them (Hillier, 2014). The study proposed that this was likely because examiners for the latter group tend to set such questions on a learning management system that was not designed for these types of questions, leading to a negative testing experience for students. Another study of 55 undergraduate students noted their frustration with connection issues, lack of motivation, scarce time, and insufficient feedback on such exams (Ocak & Karakuş, 2021). This is in contrast to another study (Jaap et al., 2021) that surveyed 119 fifth-year students at a United Kingdom medical school about their experience with switching campus-based examination to remote exam delivery during the pandemic, where few students reported technical or practical issues in completing the exam. Moreover the study noted that remote examination reduced test anxiety for some students but increased it for others.

Notwithstanding examinee frustrations with such tests, it is worth noting that another study compared the performance of 45 students on 191 proctored online exams to the performance of students who took traditional exams at a university site and found no difference in success rates (Beust et al., 2018). Moreover, an earlier study found that multiple choice questions were efficient but perceived as unfair, and that the reuse of such type of questions in computer-based examinations was controversial due to the potential for students to pass on the questions to other students in future cycles (McCoubrie, 2004). However, this was in contrast to a different study (Hertz & Chinn, 2003) that analyzed two groups of candidates taking proctored computer-based exams, with 1001 examinees for a clinical social work license and 1660 for a couples therapy license. Despite the risk of earlier examinees exposing exam questions, the study found that there was no advantage to students taking the exam in a later cycle, provided that there was a sufficiently large set of questions in the exam bank.

The format of such exams allows for providing a randomized set of questions for each examinee, which comparatively make better assessments (Santos et al., 2019). However, proper randomization requires a sufficiently large test bank. Moreover, preparing a fully randomized assessment is impractical, requiring the instructor to draft a unique question for each student. For example, a class of 20 students taking an exam with 30 questions would require drafting at least 600 questions, each assigned to no more than one student, to ensure randomness (Santos et al., 2019). Properly curating a bank of questions involves calibrating items to minimize discrimination, monitoring items for exposure, and updating the items accordingly (Parshall et al., 2002). A recent study (Fowler et al., 2022) created a simulation model to analyze if randomized question pools were impacting student performance in CBT and concluded that variances in students' exam scores were within tolerable values. The study found that even in the most unfair permutations the simulated

student's score varied negatively by only five percentage points from their expected performance.

### **Examination integrity and issues of cheating**

There is some concern that unproctored online examinations may lead to academic misconduct. According to one study, an “unproctored on-line multiple-choice exam without backtracking” revealed that students who attempted the same questions at a later round tended to score slightly higher and complete the questions quicker (Klijn et al., 2022, p. 1). Another study (Harmon & Lambrinos, 2008) compared student performance on economics examinations between 2004 and 2005 and found significant differences in student performance between the three unproctored exams and one proctored exam, leading to the conclusion that students were cheating on the unproctored exams. A meta review of the literature published between 2010 and 2021 on the topic of online exam cheating (Noorbehbahani et al., 2022) analyzed 58 publications for cheating reasons, types of cheating, cheating detection, and cheating prevention. This study concluded that lowering student cheating motivation is a very efficient strategy.

To quantify the integrity of an exam, a recent study (Murdock & Brenneman, 2020) developed a formula to calculate the integrity and reliability of quizzes with the assumption that students were comparing questions on a randomized (unproctored) test and sharing answers. The authors defined integrity as “the percentage of questions any pair of students are guaranteed not to have in common,” while reliability was defined as “the likelihood that an assessment for an entire class has an integrity” of a set value (Murdock & Brenneman, 2020, p. 93). In this study, integrity was calculated using two variables: the questions on the quiz and the desired maximum number of questions in common between any two students. Reliability, however, was determined by using a hypergeometric function that included those two variables as well as the number of students taking the quiz and the quiz bank size. The authors found that maximizing integrity would minimize reliability, therefore they proposed using the intersection of the two curves as a compromise.

While instructors encourage students to practice the subject material before an exam, many expect their examination questions to remain confidential. A survey of 340 engineering educators found that frequently reusing exam questions (but not homework assignments) was significantly correlated with observed academic misconduct in the form of disclosing exam questions to other cohorts (Gehring, 2004). While most instructors tend to reuse their exam questions every year or two, many also seek ways to obtain more questions for the exams to maintain integrity. The study reported that 86% of respondents wrote their own questions, 61% used questions from textbooks other than the one assigned for the course, 24% collected questions online, and 23% borrowed them with permission from a colleague (Gehring, 2004).

### **The benefits of practice quizzes**

Scaffolding a course with continuous assessment through a variety of summative and formative assessments has a more positive outcome on student learning than relying solely on a final exam (Paloposki et al., 2024). Summative assessments also improve information retention and reduce overconfidence (Kenney & Bailey, 2021). In the early days of adopting computer testing, a study on multiple-choice exams found that students who took practice exams with feedback performed better than those that were not given the opportunity to practice (Lee-Sammons & Wollen, 1989). While high-achieving students did not significantly improve, there was significant improvement for low-achieving students. Feedback plays a significant role in improving student performance, as a study (Naujoks et al., 2022) on two groups of undergraduate educational psychology students found that the students in the group who received individualized feedback on practice test (111 students) performed better than those in the group who did not receive feedback (201 students).

A different study on high-school and middle-school students (McDermott et al., 2014) found that frequent quizzes with feedback affected the degree to which students retained information for the exam. It also found that the format of the questions, whether multiple-choice or short answer, provided the same benefits. In contrast, a study on college students (Funk & Dickson, 2011) argued that multiple-choice exams may be inaccurate in assessing student learning because the questions can only measure the ability to recognize, but not necessarily recall, relevant information. This study also found that while the average score on multiple-choice questions was “C,” the average score on short-answer questions was far below passing. However, the authors contend that lower performance on the short-answer questions could be due to students studying with the expectation of a multiple-choice exam.

Another study on 1676 students taking four different exams in an online Organizational Behavior course found that allowing students more practice on quizzes through unlimited attempts improved exam performance by nearly 6% on all four exams (Davis et al., 2020). A similar study on 1037 final-year undergraduate nursing students (Hughes et al., 2020) found that those who regularly attempted the weekly quiz multiple times (with only the highest scoring attempt contributing to their grade) were more likely to engage during lectures. The study also found that students’ overall course grade was negatively correlated with the number of attempts, meaning underachieving students tended to repeat quizzes more in order to score higher.

A study on medical students (McNulty et al., 2015) set up a practice quiz three days before an examination and planted twelve questions from the quiz in the exam. There was no significant change in student performance from earlier exams without practice quizzes, as students performed better on four questions, the same on four, and worse on four. However, exam performance was found to be correlated with quiz performance. A different

study on medical students at a different institution (Chang & Wimmers, 2017) reached the same conclusion, but added that struggling students benefited from the practice while excelling students did not. Similarly, a study on 200 computer science students concluded that students who attempted the practice quizzes performed better, and there was a weak negative correlation between the time a student spends on a quiz and their performance on the exam (Fossati & Hashemi Tonekaboni, 2020).

A recent study on engineering students used the same calculation questions on a practice quiz and an exam, only with different numbers (Cummings, 2020). The author concluded that while the students' scores on the quiz did not correlate with their scores on the exam, participation in the practice quiz correlated with improved exam scores. Similarly, a study on the impact of providing code-writing assessments in a computer programming class showed that correctly practicing such exercises led to better performance on exams (Ahadi et al., 2016).

A major study in Germany analyzed 10,148 multiple choice responses from medical students on exams across five semesters (Appelhaus et al., 2023). As a change in policy allowed the reuse and disclosure of some of the exam questions, the Fall 2017 exam had no repeated questions on their exam, while Spring 2018 had 18.2 % repeated or disclosed questions, Fall 2018 had 29.3%, Spring 2019 had 36.8%, and Fall 2019 had 48.4%. This study considered three types of questions in the analysis of the exam: (a) new and never used before on the exam, (b) used on previous exams but not disclosed to students, and (c) used on previous exams and disclosed to students. The study's results indicated that the reused, disclosed questions lead to higher discrimination. The authors argued that formative testing with feedback increased transparency and decreased student anxiety, but necessitated a larger question bank. The study concluded that repeating and disclosing questions, while benefiting students, would lead to less integrity of examinations.

Yet another study on 130 examinees repeating the Medical Council of Canada Evaluating Examination (MCCEE) analyzed their performance on 36 reused questions out of a total of 324 questions on the exam (Wood, 2009). The study found that examinees performed better on the reused questions just as much as they did on non-reused questions, leading to the conclusion that the examinees were not advantaged by having reused questions on their exam. In contrast, a similar study on 1,629 exam questions answered by Canadian undergraduate medical education students concluded that exposing a question three or more times in a short period posed a risk to the exam's psychometric properties (Joncas et al., 2018).

Another study argues that exam transparency benefits students (Maciejewski, 2021). In this study, 89 students taking a second year differential equations course were allowed a "cheat sheet" on one test, while another test allowed students to bring in any resources. The study found that there was no significant difference in student performance between the

two tests, and that the open-resource test did not inflate grades. The authors argue that such exams allow asking more open-ended, conceptual questions, and that the grading is more focused on problem-solving instead of simply penalizing calculation errors.

As noted from this literature review, there is a noticeable gap in identifying the impact of reusing questions between assessments. While some studies have attempted to test the repetition of a few questions, there has not yet been a study on determining the statistically sufficient bank size to justify using the entire bank between assessments. There is also a gap in assessing the impact of exam integrity on student performance.

## **Methods**

### **Data collection and organization**

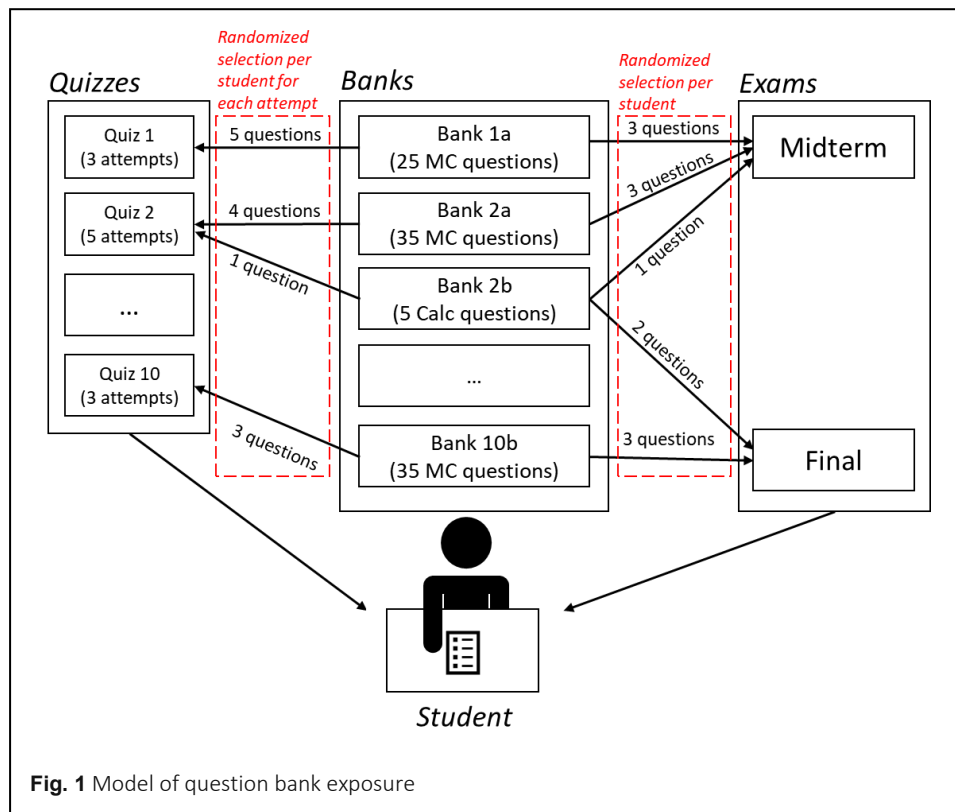
The study collected test performance data from a senior-level undergraduate Construction Management course. The authors obtained approval from the university's Internal Review Board to conduct the study according to ethical research principles, with precautions taken to ensure privacy and security of the data such as obfuscation of student identifying information and storage of on a secure cloud server. The course consisted of weekly practice quizzes, a midterm exam, and a final exam. The quizzes were considered formative assessments, while the exams were summative assessments. For each response from every student on every test, the researcher collected data from the course learning management system (Canvas) comprising the student's name, the test, the attempt number, the serial number of the question, and the student's score on that question. To maintain privacy and data security, student names were converted into serial numbers. The question serial numbers were used to determine their bank of origin. The main assumptions of this study were that students did not share quiz questions, students did not collaborate during quizzes or exams, and questions within each bank were comparable in difficulty.

The study extracted test performance data from four full semesters of a senior-level construction management course taught by one of the authors from Fall 2021 to Winter 2023. All the instructional material and examination questions were being offered for the first time, minimizing the chance that students had prior exposure to the questions beyond what they saw in the practice quizzes. The course included 10 weekly quizzes to assess students' understanding of the material, with each quiz extracting a total of five questions from two or three question banks. Extracted questions were randomized for each attempt. Each quiz allowed up to three attempts, with the highest scoring attempt counting towards the student's grade. However, quizzes related to scheduling allowed up to five attempts to account for the difficulty of the material. At the end of each quiz attempt, the learning management system would automatically grade the student's submission and display the instructor's preset feedback for wrong answers. While the midterm and final exams

extracted a fixed number of questions from the same banks used in the quizzes, students were only allowed one attempt and the only feedback they received was the total score on the assessment. To avoid the issue of different point values been allotted to the same questions on different assessments (i.e., a question worth two points on a quiz being worth three on an exam), the study normalized each performance on a question between the values of 0 and 1. Figure 1 shows a model of how question banks interface with quizzes and exams (the numbers presented are illustrative only).

For the first three semesters, there were 19 question banks in total, with the midterm and final examinations sharing some banks in common. By the fourth semester, the banks were expanded to a total of 25. Note that in Table 1 the sum of question banks used in each exam across all course offerings was not calculated (instead given an “N/A” value) because the banks themselves were being reused between semesters. However, the number of student-banks and student-questions analyzed were summed because each represents a unique student’s performance on a bank or question.

The number of “Student-Exam Banks Analyzed” in each semester was determined by the number of “Banks used in Midterm” (or Final) multiplied by the number of students. For example, in Fall 2021, 17 students were examined with seven banks in the midterm and 14 in the final. 17 students multiplied by 21 banks totaled 357 summative student-banks, representing each student’s attempt for each bank on a summative assessment. For further



**Table 1** Number of summative and formative questions and banks used in the study

Course Offering	Students	Question Banks	Banks used in Midterm	Banks used in Final	Student-Exam Banks Analyzed	Student-Exam Questions Analyzed	Student-Quiz Questions Analyzed
Fall 2021	17	19	7	14	357	1071	1921
Winter 2022	25	19	8	15	575	1575	2762
Fall 2022	13	19	8	15	299	819	1282
Winter 2023	21	25	9	23	672	1323	2394
Total	76	N/A*	N/A*	N/A*	1903	4788	8359

\*Repeated quantities were not summed

clarification, there were only 19 unique banks, as the midterm and final exams shared two banks in common. However, as the questions extracted from the banks on each student attempt were randomized, the midterm and final exam attempts on these common banks were considered as separate. The midterm contained 24 questions extracted from seven to nine banks, while the final contained 39 questions extracted from 14 to 23 banks. Adding the midterm and the final exams together, each student answered a total of 63 summative examination questions. In Fall 2021, for example, 17 students attempted 63 summative questions for a total of 1071 “Student-Exam Questions Analyzed”.

Students took 10 quizzes throughout the course, with each quiz containing five questions (except two scheduling quizzes which had three each), for a total of 48 quiz questions per student. While students could attempt a quiz up to three times (and the two scheduling quizzes allowed up to five attempts), not all students choose to use all their attempts, which resulted in the number of quizzes analyzed being less than the maximum possible 150 quiz questions per student. Of the 1903 student-banks analyzed, seven were discarded as outliers due some students being granted one extra quiz attempt when an isolated technical issue prevented users from inputting an answer, leaving a total of 1896 student-banks analyzed in total. While blank responses were graded as incorrect, these were retained as students would still benefit from the feedback after submitting their quiz.

Questions in the study were categorized into three categories: multiple-choice (MC), scheduling (Sched), and Calculation (Calc). Some distinct multiple-choice questions were also identified (DMC). These were questions with memorable scenarios or cues that would make them quite easy to recognize if repeated, such as an ailurophobic contractor trapped inside a cat shelter or a subcontractor comically named Faster Plaster Caster. The analysis included the DMC banks as part of the MC banks unless otherwise specified. Sorting-type and matching-type questions were categorized under the MC category as well.

Scheduling (Sched) questions, while similar to multiple choice questions in offering students a set of choices from a drop-down menu, were much more difficult to differentiate

between. While a student could recall the answer to a multiple-choice question from its prompt, it would be much more difficult to remember the values of durations and the corresponding correct response values in a scheduling question. These questions were also distinct from calculation questions in that their values were not randomly generated.

Calculation (Calc) questions utilized the learning management system's ability to generate several inputs to be used in a hidden formula to determine the correct answer. Not only was there randomness in pulling a question from the bank, but the question itself could have up to 200 different configurations for the input values. It would be exceedingly difficult to memorize the correct answer for a given input and get the same inputs for a repeated question of this type. However, since the method to calculate the solution remained the same, this study considered that a repeated question would lower the bank's integrity regardless of whether the question's inputs changed.

### Integrity formula derivation

Integrity was defined in this study as the portion of the questions on a summative assessment (exam) that had not previously appeared on a formative assessment (quiz). As questions were grouped in banks and randomly extracted for each attempt in a quiz or examination, each question bank would have an integrity measure for each student. The actual integrity of an exam can be determined by calculating the percentage of questions that have not been exposed to the student before in prior practice quizzes. The theoretical integrity of an examination is the expected probability that none of the questions presented in the exam had been exposed in a prior quiz that had used the same question bank.

The probability of a quiz question reappearing on an exam would be the number of questions from the bank ( $Q$ ) divided by the size of the bank ( $B$ ). Taking the complementary value of this probability (100% minus this probability) would give the probability of a question not reappearing, thus maintaining integrity. This probability is then multiplied by itself for each opportunity of exposure, which is the number of quiz attempt given to students before the exam ( $N$ ). The integrity of a summative assessment can be calculated as:

$$I = \left(1 - \frac{Q}{B}\right)^N \quad (1)$$

Where

$I$  = Integrity of summative exam; probability of no questions being previously exposed

$Q$  = Number of questions extracted from the bank for each quiz attempt

$B$  = Number of questions in the question bank

$N$  = Number of previous quizzes students attempted that used the same question bank

**Table 2** Matrix of an exam’s theoretical integrity if practice quiz allowed three attempts

Bank Size	Questions per Quiz			
	2	3	5	10
10	51%	34%	13%	0%
20	73%	61%	42%	13%
30	81%	73%	58%	30%
40	86%	79%	67%	42%
50	88%	83%	73%	51%
60	90%	86%	77%	58%
70	92%	88%	80%	63%
80	93%	89%	82%	67%
90	93%	90%	84%	70%
100	94%	91%	86%	73%

Therefore, Equation 1 can be used in the general-use case to calculate the integrity of a question bank based on the size of the bank, the number of questions being extracted on each attempt, and the prior number of attempts. Naturally, if the bank was never revealed to examinees in any prior attempts ( $N = 0$ ), then the integrity would be 100%. Likewise, if all the questions were extracted from the bank ( $Q = B$ ), then the integrity would be 0%. Using this formula, Table 2 provides a matrix of the theoretical integrity of a question bank used on a summative assessment if that bank was previously used in a formative assessment that allows up to three attempts.

While viable for most applications, Equation 1 does not account for cases where examiners wish to expose a specific number of questions in a bank but not have that count as a reduction in integrity. This might apply in the case of Calc, bonus, or essay questions. To determine such an equation, the authors applied statistical probability functions. As assessments are formed by extracting questions from a bank without replacement (i.e., the quiz or exam will not pull the same question twice on the same attempt), a hypergeometric distribution adequately defines this probability for theoretical integrity. The probability of having been exposed a specific number of questions (through quizzes) before the exam can be calculated as:

$$\frac{\left(\frac{K!}{x!(K-x)!}\right)\left(\frac{(B-K)!}{(Q-x)!(B-K-Q+x)!}\right)}{\frac{B!}{Q!(B-Q)!}} \tag{2}$$

Where

$x$  = Specific number of exposed questions on an exam

$K$  = Probable number of previously exposed questions =  $B \left(1 - \left(1 - \frac{Q}{B}\right)^N\right)$

To calculate the average probability of exposing at least  $x$  questions, the above discrete probability is calculated for each value  $i$  ranging from 1 to  $x$  and respectively multiplied by the same value of  $i$ . The sum of these values would represent the expected number of questions exposed. Dividing by the number of questions on an exam would produce the portion of questions exposed. Therefore, if the definition of integrity is expanded to allow for up to a specified number of questions to be exposed without impacting integrity, the modified integrity can be calculated as:

$$I = 1 - \sum_{i=x}^N \left( \frac{\left( \frac{K!}{i!(K-i)!} \right) \left( \frac{(B-K)!}{(Q-i)!(B-K-Q+i)!} \right) \frac{x}{Q}}{\frac{B!}{Q!(B-Q)!}} \right) \quad (3)$$

One concern with using factorials is that some software applications have calculation issues if the inputs are not whole numbers. In particular, the probable number of previously exposed questions ( $K$ ) is not necessarily a whole number, and this value is one of the inputs in the function. To get around this limitation, a Gamma distribution can be used as a substitute for a factorial of any value  $A$  where  $A! = \Gamma(A + 1)$ . The substitution results with the formula:

$$I = 1 - \sum_{i=x}^N \left( \frac{\left( \frac{\Gamma(K+1)}{i!\Gamma(K-i+1)} \right) \left( \frac{\Gamma(B-K+1)}{(Q-i)!\Gamma(B-K-Q+i+1)} \right) \frac{x}{Q}}{\frac{B!}{Q!(B-Q)!}} \right) \quad (4)$$

Therefore, Equation 3 can be used to calculate the integrity of a question bank with multiple attempts while intentionally allowing a specific number of questions to be repeated without considering such repetition as impacting integrity. Equation 4 produces the same result but can instead be used in software applications that cannot properly calculate partial factorials.

A Monte-Carlo simulation was conducted to verify the above formulas with 100,000 randomly generated samples, where the number of questions extracted from the bank for the exam ( $Q$ ) ranged from 5 to 100 in increments of 5, the number of questions in the question bank ( $B$ ) ranged from 10 to 100 in increments of 5, and the number of previous quizzes students attempted that used the same question bank ( $N$ ) varied between 1, 3, and 5. For each iteration, the simulation randomly extracted  $Q$  questions from an array numbered from 1 to  $B$  and repeated this process 100,000 ( $N$ ) times to produce a set of numbers that represented questions exposed before the exam. The simulation then extracted another  $Q$  questions from the bank array to represent the exam questions and calculated the portion of these exam questions that had been exposed. Calculating the

complement of this portion (by subtracting it from 100%) would produce the integrity ( $I$ ) of the exam.

### **Bank integrity impact on assessment authenticity**

Most of the responses were graded as either correct (1) or incorrect (0) with the exception of a few questions that could net a fractional grade such as those in a multiple-selection, matching, or sorting format. Due to the overwhelmingly polar nature of responses, where data was mostly clustered around 0 or 1 with very few values between, significant differences between non-normal distributions were analyzed using non-parametric tests. The Friedman test was used for the analysis of more than two groups of repeated measures (such as performance on quizzes vs exams), with the Wilcoxon Signed Ranks test for post-hoc testing of two groups. Kruskal-Wallis was used to analyze more than two groups of non-repeated measures (such as comparing theoretical integrity to actual integrity), with the Brunner-Munzel test used for post-hoc testing of two groups. This post-hoc test was preferred to the more traditional Mann-Whitney U-test as per recent research (Fagerland & Sandvik, 2009; Karch, 2021). For all these statistical tests, the study assumed the results were significant at a p-value less than or equal to 0.05.

While one of the aims of this study was the derivation and verification of the above formula, the study also aimed to determine if reusing questions in practice quizzes can benefit students without compromising summative examinations. The study approached this question by analyzing two key aspects: whether the performance on an exam was influenced by exam integrity and whether repetition benefited student learning.

To investigate the influence of integrity on exam performance, the study first analyzed the correlation between integrity and exam performance using the Pearson coefficient. Each student-bank's integrity was compared to its highest score on the quiz, score on the exam, the difference between the exam score and the highest score on the quiz, and the difference between the exam score and the average score of all quiz attempts. The study then compared the performance between quizzes and exams using the Friedman test and the Wilcoxon Signed Ranks tests. This was followed by a Kruskal-Wallis test between high- and low-integrity scores for each bank of the Midterm and Final exams to identify if the significance was occurring consistently across all banks or between an isolated set, starting with a cut-off point of 70% integrity (to define the boundary between high- and low-integrity) then testing if the results would be affected if the boundary were to be changed to 50%, 60%, or 80% integrity.

### **Gauging the benefits of question repetition**

To investigate the impact of repetition on student learning, the study tested whether more attempted questions lead to better performance by subdividing student attempts into five

categories and conducting a Kruskal-Wallis test followed by a Brunner-Munzel test for significant pairs. A student-bank with less than three attempts was labeled “Very Low,” three to five was “Low,” six to eight was “Medium,” nine to eleven was “High,” and twelve or more was “Very High”. The impact of repetition was also investigated by analyzing the exam performance of students who had been exposed to a question before to those who had not using a Brunner-Munzel test. This was further analyzed by subdividing the students in groups based on number of times a question was repeatedly exposed and comparing exam performance using a Kruskal-Wallis test followed by a Brunner-Munzel test for significant pairs.

Lastly, the study investigated whether the improvement of student performance on quizzes lead to improved performance on exams. The student-banks were divided into two groups (Groups 1 and 2) based on whether the student-bank’s highest score on the quiz achieved a set passing grade or not. The passing grade was set at two different levels, one at 60% to represent a minimum acceptable degree of performance, and again at 80% to represent a higher expectation of the level of understanding. The groups were further subdivided into groups based on the number of quizzes attempted. This is not to be confused with the previous analysis on question repetitions, as this analysis focused on quiz attempts, where a student who only attempted a quiz once would have had no chance of any repeated questions (this is expected to be the case where a student was satisfied with their quiz performance on their first try). The test then conducted a Brunner-Munzel test on the average student-bank exam score between these two groups.

## **Results and discussion**

The below results address the three main research questions presented in this study. With a formula derived to address the first question (“how many questions should a bank have to achieve a desired integrity?”), the study tested the formula’s accuracy and compared it to the actual integrity of the exams within the study’s dataset. The second question (“does exposing questions ahead of an assessment reduce its authenticity?”) was investigated through examining the correlation between integrity and performance, comparing the performance on quizzes versus exams, and testing for significance between high- and low-integrity exams. The third research question (“does allowing repeated exposure to questions benefit student learning?”) was addressed through testing of performance versus the number of attempted questions, comparing performance versus exposed questions, comparing performance versus multiple question exposure, and testing the effect of improvements in quizzes on exam performance.

### Formula accuracy

When analyzing the results of the Monte-Carlo simulation, the theoretical integrity formulas tended to produce inaccurate values (deviations greater than 5%) if the number of questions on the assessment ( $Q$ ) exceeded 75% of the bank size ( $B$ ), and this error became more pronounced with more quiz attempts ( $E$ ). However, when all the results of the simulation were compared against the respective values using theoretical integrity formulas, the error rate was found to be less than 1% on average. When excluding values where  $Q$  exceeds 75% of  $B$ , the error rate dropped to 0.2%. Because integrity would be quite low ( $< 25\%$ ) in cases where  $Q$  exceeds 75% of  $B$ , examiners seeking to set a test with reasonable integrity would anyway avoid such high overlap. Accordingly, these results indicate that the derived formulas are accurate for most applicable purposes. These results also align with the verification methods used by Murdock and Brenneman (2020).

### Actual integrity higher than theoretical

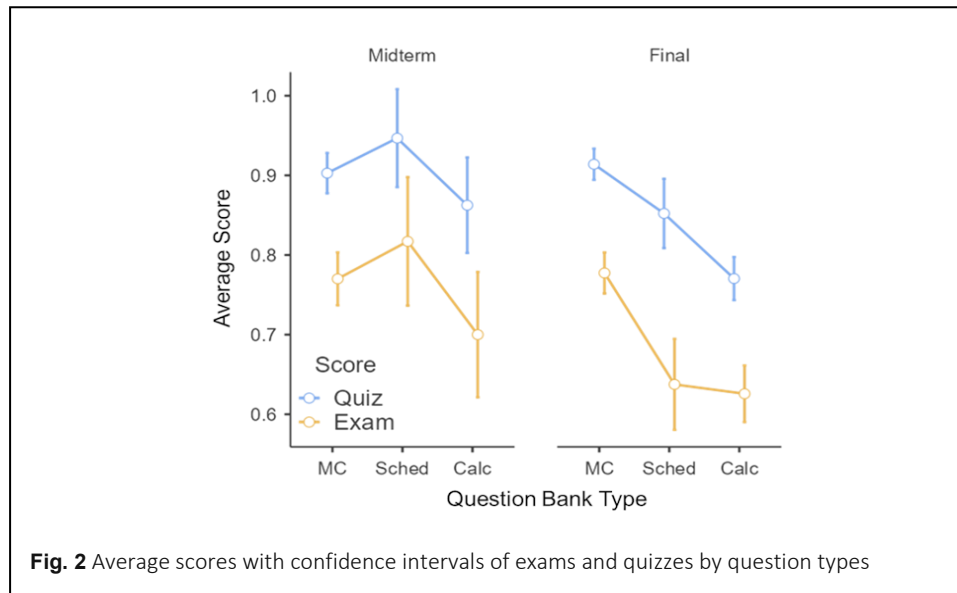
To further validate the formula, the study compared the actual number of exposed questions to the student on exams with the theoretical integrity using the derived formula. One of the early significant findings of this study was that actual integrity tended to be higher than theoretical integrity. For each bank of questions, actual integrity was calculated by taking the portion of exam questions for each student that were not previously exposed in a quiz and calculating the average for the bank, while the theoretical integrity was calculated using the formula explained earlier. Using the Wilcoxon Signed Ranks test to compare the actual and theoretical integrity of the 1896 student banks showed that actual integrity was 11% higher than theoretical integrity at a p-value of 0.007. The likely explanation is that because students did not use all their attempts, their exposure to questions was lower.

### Integrity not correlated with performance

The study used the Pearson coefficient to calculate the correlation of integrity with student performance. The results in Table 3 indicate that there was no significant correlation between integrity and any of the four measures of performance, even when the banks were subdivided by type. A Pearson coefficient less than 0.3 is generally considered a weak correlation.

**Table 3** Pearson correlation coefficient (R) values of integrity with student performance, subdivided by bank type

Correlation	All	MC	Sched	Calc
Integrity with "Highest Quiz Score"	0.057	-0.018	-0.121	-0.010
Integrity with "Exam Score"	-0.070	-0.198	-0.209	-0.065
Integrity with "Exam – Highest Quiz Score"	-0.101	-0.150	-0.128	-0.053
Integrity with "Exam – Average Quiz Score"	-0.211	-0.211	-0.167	-0.214



### Performance on quizzes was better than exams

Another significant finding was that students performed better on quizzes than exams, even though the exams reused the same banks from the quizzes. Using the Wilcoxon Signed Ranks test to compare the performance of 1896 student-banks in their quizzes against the performance in the exams showed a very high significance ( $p < 0.001$ ) with quizzes outperforming exams by nearly 15% for all types of banks (see Figure 2). One possible explanation for this could be students' ability to reattempt quizzes immediately, while the exams were scheduled many weeks later. However, the authors do not have sufficient data to conclusively prove this conjecture. This finding does, however, support the theory that quizzes by themselves do not necessarily inflate exam grades (but practice quizzes do help if students learn from their mistakes, as discussed in the later sections of this paper).

### No significance in performance between high- and low-integrity banks

In contrast to the above, when the 1896 student-banks were subdivided into high-integrity (integrity at least 70%) and low-integrity groups (integrity less than 70%), the Brunner-Munzel test comparing the exam scores between these two groups showed a significant difference ( $p < 0.001$ ), with low-integrity exams scoring about 10% higher than high-integrity ones. To determine the cause of this significance, the student-banks were further subdivided by Exam Type and then by Bank Type (see Table 4). The Final Exam showed significant differences for all types, with Multiple Choice banks in the low-integrity group scoring higher by nearly 10% compared to those in the high-integrity group (Table 5). Likewise, low-integrity Scheduling and Calculation banks scored higher by nearly 29% and 14%, respectively, compared their higher-integrity banks.

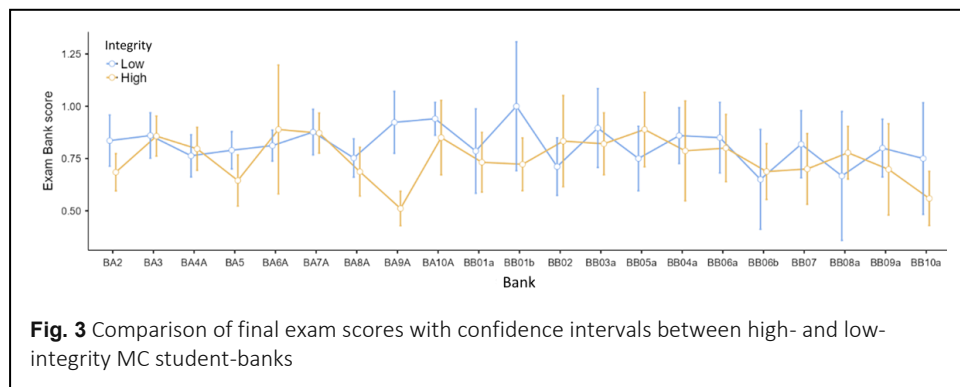
**Table 4** Subdivision of student-banks

Exam	Type	Integrity	N
Midterm	MC	Low	219
		High	230
	Sched	Low	11
		High	65
		Calc	49
Final	MC	Low	385
		High	362
	Sched	Low	25
		High	127
		Calc	307
		High	85

**Table 5** P-values of Brunner-Munzel tests comparing high- and low-integrity banks, subdivided by bank type

Type	Midterm Exam	Final Exam
All	0.166	< 0.001
MC	0.010	0.004
Sched	0.184	< 0.001
Calc	0.372	0.029

This analysis revealed that the significance was indeed caused by an isolated set of banks. Every time a significance was discovered, the significant bank was removed from the set, and the Kruskal-Wallis test was conducted again. After only seven banks were systematically excluded, no significance was found between low-integrity and high-integrity exam scores within any of the remaining banks. Those seven banks consisted of four multiple-choice banks (BA1B, BA2, BA5, and BA9A), two scheduling banks (BA8B, BB10B), and one calculation bank (BA7B). Figure 3 compares the final exam score of each multiple-choice high-integrity student-bank against low-integrity multiple-choice student-banks, where the deviation of banks BA2, BA5, and BA9A can be seen (BA1B is missing as it was used in the midterm but not the final exam).



The authors identified all these banks as containing sets of questions more challenging than banks covering similar topics. For example, Banks BA1A and BA1B both cover the topic of evaluating bids, but BA1B consists of questions where a “best-value” analysis needs to be conducted on a set of bidders with an outlier that needs to be eliminated first before setting the lowest price for comparison. With this identified set of seven challenging banks excluded, the Kruskal-Wallis test was repeated using integrity values of 50%, 60%, and 80% for cut-off points between high and low integrity, which also resulted with non-significant p-values of 0.087, 0.360, and 0.427, respectively.

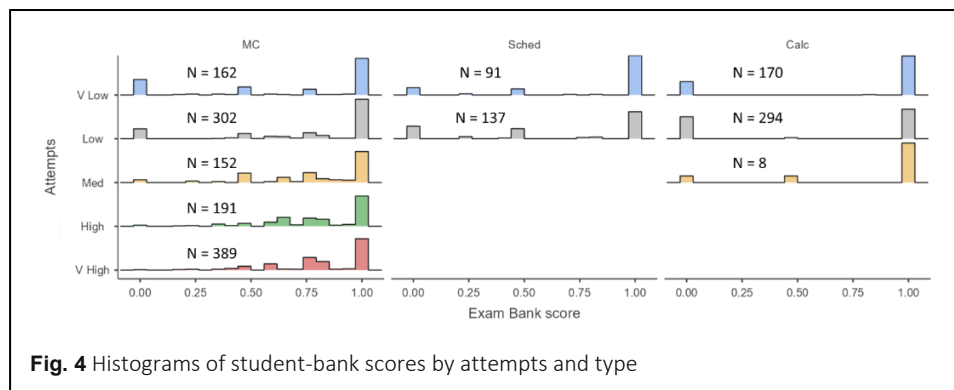
Therefore, while lower integrity led to higher performance on tricky or challenging banks, the data showed that there was no significance in performance between high and low integrity banks when controlling for challenging banks, regardless of the cut-off point between high and low integrity. This would imply that sharing banks between quizzes and exams allows students to learn from their mistakes on challenging questions without giving an unfair advantage on the rest of the summative assessment.

**More attempted questions do not necessarily lead to better performance**

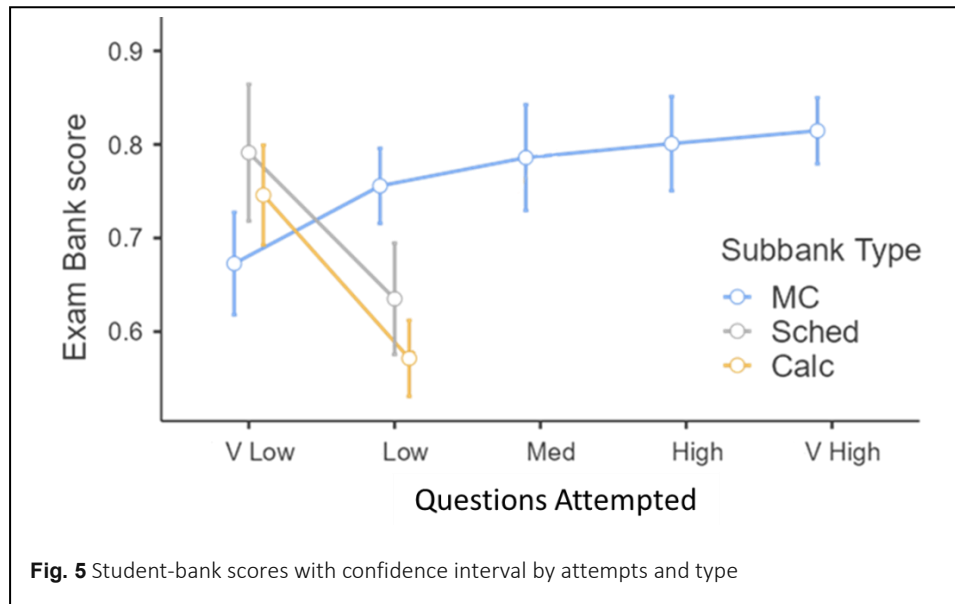
The study analyzed performance based on the number of questions attempted from a given bank. A student-bank with less than three attempts was labeled “Very Low,” three to five was “Low,” six to eight was “Medium,” nine to eleven was “High,” and twelve or more was “Very High” (visualized in Figure 4).

The quizzes would extract between one to five questions from the banks, and students usually had up to three attempts on quizzes, so the range of attempted questions from a bank was from one to fifteen. However, quizzes were designed to select five questions from multiple choice banks but only one or two from scheduling and calculation banks, which led to limited opportunities to attempt more than five questions from the latter two types.

The analysis detected significant differences (p-value = 0.005) in the average of the banks’ scores on the final exam between students who attempted twelve or more multiple-choice questions in a bank on their quizzes (83%) and those who attempted less than twelve (75%).



**Fig. 4** Histograms of student-bank scores by attempts and type



This significance was still detected to the same extent when controlling for challenging banks. However, the average of the bank score on the final exam among students who attempted three or more scheduling questions (59%) was lower than those who attempted less than three (72%). The same discrepancy was detected in calculation banks. Figure 5 shows how performance oddly dropped on scheduling and calculation student-banks with more attempts.

On the other hand, the possibility of repeating the same question over multiple quiz attempts posed a concern for the above analysis. Students who get the same question over several quiz attempts would do better on the quizzes but would have been exposed to a smaller portion of the bank, leaving them with a higher possibility of getting a question on the exam that they have not practiced before. To control for this, the study repeated the above analysis but limited the dataset to the number of unique questions attempted (i.e., the number of questions across all quiz attempts less any that had been repeated between attempts). When challenging questions were excluded, no significance was found across various separation points between the number of attempts and exam performance (see Table 6). This means that practicing a bigger variety of questions from a bank does not necessarily lead to better performance on an exam.

**Table 6** P-values comparing student-bank final exam scores by unique question attempts, subdivided by type

Grouping by Unique Attempts	All	MC	Sched	Calc
< 3 vs 3 or more	0.288	0.734	0.554	0.723
< 5 vs 5 or more	0.194	0.959	0.231	-
< 7 vs 7 or more	0.869	0.129	-	-
< 9 vs 9 or more	0.516	0.099	-	-
< 12 vs 12 or more	0.566	0.362	-	-

**Table 7** Students' performance on exposed and unexposed questions

Type	Exposed	Size	Mean Score
MC	Yes	998	90.6%
	No	1627	75.9%
DMC	Yes	442	90.5%
	No	733	73.0%
Sched	Yes	397	62.2%
	No	216	52.8%
Calc	Yes	95	88.4%
	No	258	70.5%

### Students perform better on exposed exam questions

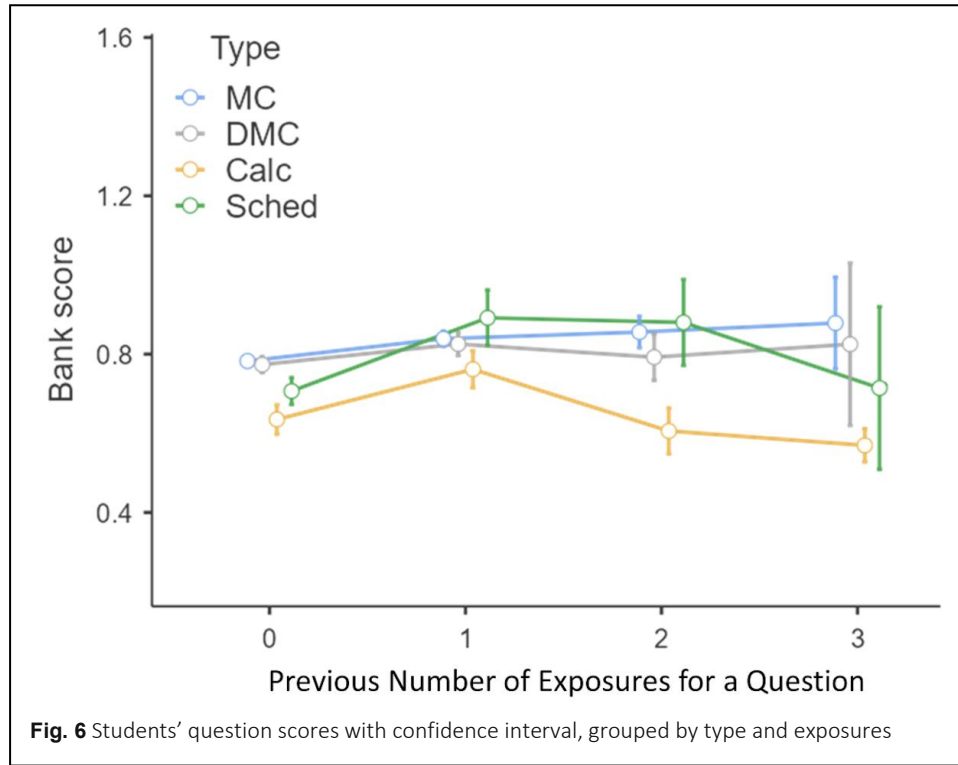
Unsurprisingly, the analysis showed that students who got an exam question that they had seen before on a quiz tended to do significantly better ( $p$ -value  $< 0.001$  for all types), as shown in Table 7. However, when removing the set of challenging questions, the calculation questions no longer show significance. The explanation for this is that calculation questions use a set of randomly generated inputs that are very unlikely to be repeated on a subsequent generation of the problem.

### Exposing a question more than once does not improve exam performance

Taking the previous test a step further, the study analyzed the performance of students on questions that were exposed more than once in quizzes. The analysis showed that MC and DMC questions had a significant difference in a student's performance between questions that were not previously seen in a quiz (0 exposed) and those that had been exposed once or twice ( $p$ -value  $< 0.001$  for both cases). However, when comparing the performance of 1,627 unexposed MC questions to 22 MC questions that were exposed thrice, the  $p$ -value was 0.10. Similarly, for Sched and Calc questions, significance ( $p$ -value of 0.003 and 0.010 respectively) was only detected between unexposed questions and those that were exposed only once. There was no significance between the 258 unexposed scheduling questions and the 25 questions exposed twice. Likewise, there was no significance between the 216 unexposed calculation questions and the 89 calculation questions exposed twice nor the 169 calculation questions exposed thrice (Figure 6). This may indicate that any exposure to MC and DMC questions would significantly improve performance. Furthermore, the lack of significance for more than one exposure to scheduling and calculation questions may be due to students repeatedly attempting a question without properly understanding the process to properly solving it.

### Improvement in quizzes leads to better exam performance

To explain the lack of significance in the previous test for repeated question exposure, the study then conducted an analysis to compare students' performance on exam questions by



grouping them according to their performance on quiz questions. The analysis showed that students who attempted quizzes at least twice and did well enough ( $\geq 60\%$ ) or even quite well ( $\geq 80\%$ ) did significantly better on exams compared to other students. On the other hand, the analysis did not detect significance between groups for only one quiz attempt. Considering that more attempts would lead to lower integrity, the authors conducted another test limited to only student-banks that did not have any exposed questions (100% integrity). The results showed that students who attempted quizzes at least three times and did well enough ( $\geq 60\%$ ) or even quite well ( $\geq 80\%$ ) did significantly better on exams compared to other students (Table 8).

**Table 8** Full-integrity student-bank performance based on the number of attempts and passing grade

Passing Grade	Quiz Attempts	Size		Mean Exam Score		P-value
		Group 1	Group 2	Group 1	Group 2	
60%	All	588	90	0.672	0.586	0.072
60%	1	156	8	0.719	0.762	0.777
60%	2	133	6	0.683	0.917	0.092
60%	3	213	26	0.650	0.365	0.010
80%	All	558	120	0.677	0.587	0.010
80%	1	152	12	0.722	0.708	0.770
80%	2	130	9	0.689	0.752	0.785
80%	3	204	35	0.643	0.479	0.028

However, no significance was detected when the challenging student-banks were excluded from the analysis. This indicates that students who persevered at attempting challenging questions until they answered them correctly by the third try significantly outperformed their peers who ultimately were unable to correctly answer those questions. One interpretation of this result is that students who did not grasp the concept after three attempts tended to underperform on exams. An explanation for this is that students were aware of the limited number of attempts, leading some students to rethink their approach and review the topic after two failed attempts, while other students pressed on with their last attempt without revisiting their strategy.

## Conclusions

This study aimed to numerically analyze the benefits and risks of sharing question banks between formative and summative assessments. While ultimately keeping two separate banks for formative and summative assessments would be the ideal approach to maintaining the integrity of summative assessments, this would require substantial preparation from the instructor. This study has shown that sharing the banks between these two types of assessments can provide significant learning benefits without significantly influencing the summative assessment. The study also provided a simple formula for predicting the integrity of a question bank based on the size of the question bank, the number of questions extracted from the bank, and the number of exposures before the summative assessment.

Examiners can use this formula to determine the required question bank size for a desired integrity in the case where the summative and formative assessments share the same question banks. Actual integrity was found to be higher than theoretical by 11% when students were allowed multiple attempts and could opt not to use all attempts. However, while this difference would increase with more quiz attempts, the integrity would exponentially decrease. Examiners should therefore be wary that allowing more attempts would require larger question banks to maintain the same integrity.

The findings of this study have several practical applications. The most prominent benefit is that it provides educators with justifications to experiment with new ways to help students practice the material they are learning. It also provides studies on pre-exposing questions (such as McNulty et al., 2015) with a measure for the impact on the exam's integrity. These findings can also aid examiners in determining if a CBT exam is using sufficiently sized question banks if a failing student is allowed to retake the exam after receiving feedback on it.

However, one major caveat of this study can be summarized in an alternative version of a popular saying: "Practice makes permanent." The study found that students do not necessarily benefit from just having some questions exposed in practice quizzes, even with

automatic feedback. While the analysis unsurprisingly revealed that students perform better on exposed questions, it also revealed that exposing a question more than once does not significantly impact a student's performance on exams. Moreover, the analysis showed that students need to improve and learn from mistakes on quizzes for that to translate into improved performance on summative assessments; students who did not adapt their approach continued to incorrectly answer those questions on summative assessments. One of the main assumptions of this study was that students would receive feedback on quizzes, but not exams. Examiners should be wary that a lack of sufficient feedback on practice quizzes might not achieve the same results in learning improvement.

This study has vast potential for further future research. It could be repeated on larger class sizes, over larger periods, or on other course topics. One particular area of potential future study could be on courses that rely almost entirely on calculation-type questions, such as engineering or math courses. However, some types of problems might be potentially difficult to program proper calculations for the solutions within the limitations of the learning management system.

#### **Abbreviations**

Calc: Calculation-type question banks; CBT: Computer based testing; DMC: Distinct Multiple-Choice question banks; MC: Multiple-Choice question banks; MCCEE: Medical Council of Canada Evaluating Examination; Sched: Scheduling-type question banks.

#### **Authors' contributions**

The first author was contributed to the conceptualization, literature review, data analysis, and writing of this manuscript. The second author contributed with literature review, data organization, manuscript review and editing.

#### **Authors' information**

R.J. joined Eastern Michigan University in 2021 as an Assistant Professor, where he teaches courses in Construction Management as well as Civil Engineering. I.I. is an independent researcher based in Sacramento, California.

#### **Funding**

This research received no external funding.

#### **Availability of data and materials**

Not applicable.

#### **Declarations**

#### **Competing interests**

The authors declare that they have no competing interests.

#### **Author details**

<sup>1</sup> Raed Jarrah, Ph.D. Eastern Michigan University, USA. <https://orcid.org/0000-0001-8955-4111>

<sup>2</sup> Ibrahim Itani, Ph.D. Independent Researcher, California, USA. <https://orcid.org/0000-0002-8709-5799>

Received: 9 July 2024 Accepted: 7 March 2025

Published online: 1 January 2026 (Online First: 5 August 2025)

#### **References**

Ahadi, A., Lister, R., & Vihavainen, A. (2016). On the number of attempts students made on some online programming exercises during semester and their subsequent performance on final exam questions. In A. Clear, E. Cuadros-

- Vargas, J. Carter & Y. Tupac (Eds.), *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 218–223). ACM. <https://doi.org/10.1145/2899415.2899452>
- Appelhaus, S., Werner, S., Grosse, P., & Kämmer, J. E. (2023). Feedback, fairness, and validity: Effects of disclosing and reusing multiple-choice questions in medical schools. *Medical Education Online*, 28(1), 2143298. <https://doi.org/10.1080/10872981.2022.2143298>
- Beust, P., Duchatelle, I., & Cauchard, V. (2018). Exams taken at the student's home. *EADTU. Online, Open and Flexible Higher Education Conference*. EADTU. <https://hal.science/hal-02129191>
- Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLOS ONE*, 10(12). <https://doi.org/10.1371/journal.pone.0143616>
- Chang, E. K., & Wimmers, P. F. (2017). Effect of repeated/spaced formative assessments on medical school final exam performance. *Health Professions Education*, 3(1), 32–37. <https://doi.org/10.1016/j.hpe.2016.08.001>
- Cummings, A. T. (2020). Correlation of student participation in practice exams and actual exam performance. *ASEE North Midwest Section Annual Conference 2020 Poster Publications*, 10. [https://openprairie.sdstate.edu/asee\\_nmws\\_2020\\_pubs/18/](https://openprairie.sdstate.edu/asee_nmws_2020_pubs/18/)
- Davis, M., Duryee, L., Schilling, A., Loar, E., & Hammond, H. (2020). Examining the impact of multiple practice quiz attempts on student exam performance. *Journal of Educators Online*, 17(2). <https://eric.ed.gov/?id=EJ1268917>
- Fagerland, M. W., & Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine*, 28(10), 1487–1497. <https://doi.org/10.1002/sim.3561>
- Fossati, D., & Hashemi Tonekaboni, N. (2020). Practice exams and student performance in introductory programming. In J. Zhang, M. Sherriff, S. Heckman, P. Cutter & A. Monge (Eds.), *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 1362–1362). ACM. <https://doi.org/10.1145/3328778.3372676>
- Fowler, M., Smith, D. H., Emeka, C., West, M., & Zilles, C. (2022). Are we fair?: Quantifying score impacts of computer science exams with randomized question pools. In L. Merkle, M. Doyle, J. Sheard, L.-K. Soh & B. Dorn (Eds.), *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education* (pp. 647–653). ACM. <https://doi.org/10.1145/3478431.3499388>
- Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, 38(4), 273–277. <https://doi.org/10.1177/0098628311421329>
- Gehringer, E. F. (2004). Reuse of homework and test questions: When, why, and how to maintain security? In *Proceedings of 34th Annual Frontiers in Education* (pp. 1033–1038). IEEE. <https://doi.org/10.1109/FIE.2004.1408702>
- Harmon, O. R., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *The Journal of Economic Education*, 39(2), 116–125. <https://doi.org/10.3200/JECE.39.2.116-125>
- Hertz, N. R., & Chinn, R. N. (2003). Effects of item exposure for conventional examinations in a continuous testing environment. *Annual Meeting of the National Council on Measurement in Education*. <https://files.eric.ed.gov/fulltext/ED476422.pdf>
- Hillier, M. (2014). The very idea of e-Exams: Student (pre)conceptions. In B. Hegarty, J. McDonald & S.-K. Loke (Eds.), *Rhetoric and Reality: Critical perspectives on educational technology. Proceedings ASCILITE Dunedin 2014* (pp. 77–88). Australasian Society for Computers in Learning in Tertiary Education.
- Hughes, M., Salamonson, Y., & Metcalfe, L. (2020). Student engagement using multiple-attempt 'Weekly Participation Task' quizzes with undergraduate nursing students. *Nurse Education in Practice*, 46, 102803. <https://doi.org/10.1016/j.nepr.2020.102803>
- Ilgaz, H., & Afacan Adanir, G. (2020). Providing online exams for online learners: Does it really matter for them? *Education and Information Technologies*, 25(2), 1255–1269. <https://doi.org/10.1007/s10639-019-10020-6>
- Jaap, A., Dewar, A., Duncan, C., Fairhurst, K., Hope, D., & Kluth, D. (2021). Effect of remote online exam delivery on student experience and performance in applied knowledge tests. *BMC Medical Education*, 21(1), 86. <https://doi.org/10.1186/s12909-021-02521-1>
- Joncas, S. X., St-Onge, C., Bourque, S., & Farand, P. (2018). Re-using questions in classroom-based assessment: An exploratory study at the undergraduate medical education level. *Perspectives on Medical Education*, 7(6), 373–378. <https://doi.org/10.1007/s40037-018-0482-1>
- Karch, J. D. (2021). Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, 4(2). <https://doi.org/10.1177/2515245921999602>
- Kenney, K. L., & Bailey, H. (2021). Low-stakes quizzes improve learning and reduce overconfidence in college students. *Journal of the Scholarship of Teaching and Learning*, 21(2). <https://doi.org/10.14434/josotl.v21i2.28650>
- Klijn, F., Mdaghri Alaoui, M., & Vorsatz, M. (2022). Academic integrity in on-line exams: Evidence from a randomized field experiment. *Journal of Economic Psychology*, 93, 102555. <https://doi.org/10.1016/j.joep.2022.102555>
- Lee-Sammons, W. H., & Wollen, K. A. (1989). Computerized practice tests and effects on in-class exams. *Behavior Research Methods, Instruments, & Computers*, 21(2), 189–194. <https://doi.org/10.3758/BF03205581>
- Maciejewski, W. (2021). Let your students cheat on exams. *PRIMUM*, 31(6), 685–697. <https://doi.org/10.1080/10511970.2019.1705450>
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, 26(8), 709–712. <https://doi.org/10.1080/01421590400013495>

- McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3–21. <https://doi.org/10.1037/xap0000004>
- McNulty, J. A., Espiritu, B. R., Hoyt, A. E., Ensminger, D. C., & Chandrasekhar, A. J. (2015). Associations between formative practice quizzes and summative examination outcomes in a medical anatomy course: Formative practice quizzes and summative examinations. *Anatomical Sciences Education*, 8(1), 37–44. <https://doi.org/10.1002/ase.1442>
- Murdock, M., & Brenneman, M. (2020). Designing tests from question pools with efficiency, reliability, and integrity. *Journal of Contemporary Chiropractic*, 3. <https://journal.parker.edu/article/78087-designing-tests-from-question-pools-with-efficiency-reliability-and-integrity>
- Naujoks, N., Harder, B., & Händel, M. (2022). Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. *Metacognition and Learning*, 17(2), 479–498. <https://doi.org/10.1007/s11409-022-09295-x>
- Noorbehbahani, F., Mohammadi, A., & Aminazadeh, M. (2022). A systematic review of research on cheating in online exams from 2010 to 2021. *Education and Information Technologies*, 27(6), 8413–8460. <https://doi.org/10.1007/s10639-022-10927-7>
- Ocak, G., & Karakuş, G. (2021). Undergraduate students' views of and difficulties in online exams during the COVID-19 pandemic. *Themes in eLearning*, 14, 13–30. <https://files.eric.ed.gov/fulltext/EJ1305321.pdf>
- Paloposki, T., Virtanen, V., & Clavert, M. (2024). From a final exam to continuous assessment on a large Bachelor level engineering course. *European Journal of Engineering Education*, 50(1), 164–177. <https://doi.org/10.1080/03043797.2024.2334728>
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer New York. <https://doi.org/10.1007/978-1-4613-0083-0>
- Santos, M. R., Richman, V., & Jiang, J. (2019). Online teaching: A study for the effectiveness of randomized exams. *Journal of Instructional Pedagogies*, 22. <https://files.eric.ed.gov/fulltext/EJ1216823.pdf>
- Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education*, 14(4), 465–473. <https://doi.org/10.1007/s10459-008-9129-z>

### Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Research and Practice in Technology Enhanced Learning (RPTEL)**  
is an open-access journal and free of publication fee.