


RESEARCH ARTICLE

Open Access



# Repeat individualized assessment using isomorphic questions: a novel approach to increase peer discussion and learning

Russell Millar and Sathiamoorthy Manoharan\* 

\*Correspondence:  
s.manoharan@auckland.ac.nz  
University of Auckland,  
Auckland, New Zealand

## Abstract

It is demonstrated that the fully automatic generation of isomorphic questions allows for both repeat assessment, and for this assessment to be individualized. While this does require a substantial up-front effort, once prepared assessments can be reproduced with relative ease and with a near-zero probability of students receiving the same question a second time. We show the effectiveness of this approach using survey and performance data obtained from large year 2 and year 3 computer science classes. A greater than 10-fold increase in online peer discussion was observed, compared to the previous year. Contents analysis of the surveys showed that repeat testing was generally regarded favourably. Quantitative analyses found that prior homework did little to improve initial test performance, but was of vital importance in studying for the follow-up isomorphic test. Moreover, the performance gain on the isomorphic questions was the same for all students regardless of their overall ability, and was retained in the final exam.

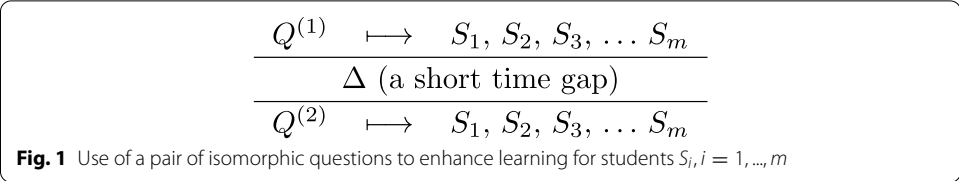
**Keywords:** Automated assessment, Online discussion, Peer learning, Student assessment, Student experience

## Introduction

Two questions are said to be isomorphic if they are both based on the same concept and test the same learning outcome. A trivial example would be the pair of questions given by “what is the sum of 2 and 8?” and “what is the sum of 5 and 7?”

The concept of using isomorphic questions to reinforce student learning is not new (Singh 2008a, b; Zingaro and Porter 2015; Kjolsing and Einde 2016). Typically, an instructor would require students to attempt a pair of isomorphic questions, the second of which will be done following a short time gap after completing the first. In the *Peer Instruction* model originally proposed by Crouch and Crouch and Mazur (2001), the short time gap is used by students to discuss in small groups the first of the questions in the pair. Figure 1 illustrates this for a repeat assessment, where  $Q^{(1)}$  and  $Q^{(2)}$  denote the first and repeat versions of the question respectively.

Use of repeat assessments by definition requires an instructor to prepare two sets of assessments. One way to reduce the additional preparation time is to use, where

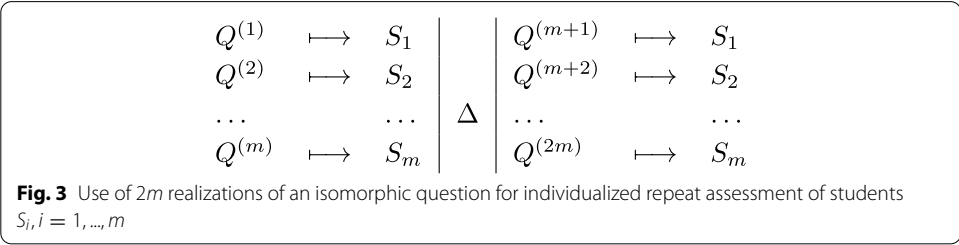
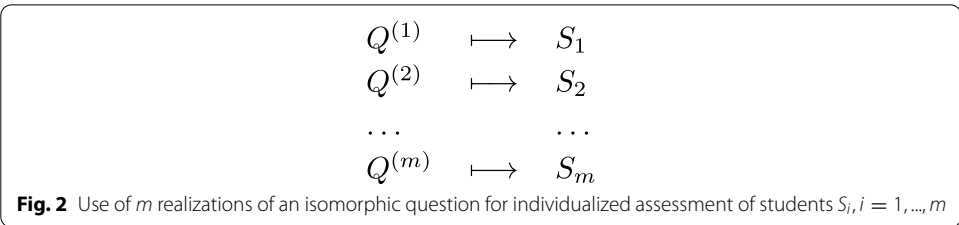


possible, forms of isomorphic questions where the difference between the question pair is only in the parameters the questions use. For example, the trivial summation questions used above are two realizations of the parameterized question “what is the sum of \$a and \$b?” where the two parameters \$a and \$b are replaced by suitable random integers at question delivery time. Here, we refer to such forms of isomorphic questions as parameter-varying isomorphic, notwithstanding that in general the varying “parameters” could be of arbitrary type.

Isomorphic questions also facilitate the application of individualized assessments (as implemented in frameworks such as Coderunner Lobb and Harlow 2016, Dividni Manoharan 2019, and OASIS Smaill 2005), whereby all students typically get different isomorphic variants of the same set of questions. Figure 2 illustrates this. We remark that where an individualized assessment is carried out, an instructor may be able to re-purpose the individualized assessment to include a repeat at little additional cost. Figure 3 illustrates this procedure.

In this paper, we investigate the consequences of introducing repeat individualized assessment into two large computer science courses. In each course we set up two individualized isomorphic midterm tests, with a one-week gap after the first test to prepare for the repeat test. In addition, in one course some of the isomorphic questions had previously been used in an assignment, and in both courses a small subset of the test questions was repeated a second time for use in the final exams.

The effect of introducing repeat individualized assessment is assessed by comparison with the same courses from the previous year. This includes assessment of the change in the use of online peer discussion, and of the overall final grade distributions.



Moreover, we perform a detailed statistical analysis of the improvement in performance on the repeat test, and examine whether the improvement depends on overall student ability or having previously seen the question in an assignment. In addition, we examine whether the improvement is maintained in the final exam.

The rest of this paper is organized as follows. Section 2 discusses the relevant prior research. This includes the use of isomorphic assessments as well as individualized assessments. Section 3 describes the methodology we apply to conducting repeat assessments as well as a rationale for doing so. Section 4 evaluates the use of repeat assessments using student feedback, and Section 5 provides a statistical analysis of student performance data. Section 6 discusses some of the challenges of individualized assessment. The final section concludes the paper.

### Background and related work

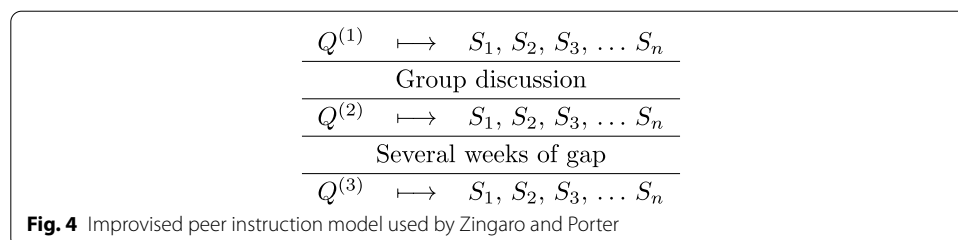
Related work in this context falls under two categories. The first category deals with the use of isomorphic questions for enhancing student learning. The second category deals with individualized or randomized assessment with the goals of providing exercises (assessed or not), reducing incidents of cheating, and encouraging independent thinking. The two categories have a common thread of using isomorphic questions. The focus of this paper is the amalgamation of the two categories using this common thread.

### Repeat isomorphic questions to enhance student learning

Zingaro and Porter track student learning through isomorphic questions in-class all the way to the exam (Zingaro and Porter 2015, 2014). The in-class questions are designed to follow the *Peer Instruction* model originally proposed by Crouch and Mazur (2001): students answer a question  $Q^{(1)}$ , discuss in small groups the question and its concepts, and then answer an isomorphic question  $Q^{(2)}$ . Zingaro and Porter add to this model another isomorphic question  $Q^{(3)}$  which is posed in the exam several weeks after the students had answered questions  $Q^{(1)}$  and  $Q^{(2)}$ . Figure 4 illustrates this model.

Their results show that

- 1 The students who scored correctly in  $Q^{(1)}$  and  $Q^{(2)}$  had the best chance of scoring  $Q^{(3)}$  correctly in the exam.



- 2 The students who did not score  $Q^{(1)}$  correctly but did score  $Q^{(2)}$  correctly had a better chance of scoring  $Q^{(3)}$  correctly in the exam than those who did not answer  $Q^{(2)}$  correctly, though not as good a chance as those who also scored  $Q^{(1)}$  correctly.

Zingaro and Porter conclude that those who learn from peer instruction in-class are likely to retain the concepts they learnt and be able to answer isomorphic questions later on in their exam.

Kjolsing and Eide report on a study similar to that of Zingaro and Porter carried out in a small engineering statics class (Kjolsing and Eide 2016). They followed the classic peer instruction model where  $Q^{(2)}$  was administered after a group discussion following  $Q^{(1)}$ . As expected, they too find that the process improved student learning. They also state that students' pre-class preparation did not affect the learning gains statistically.

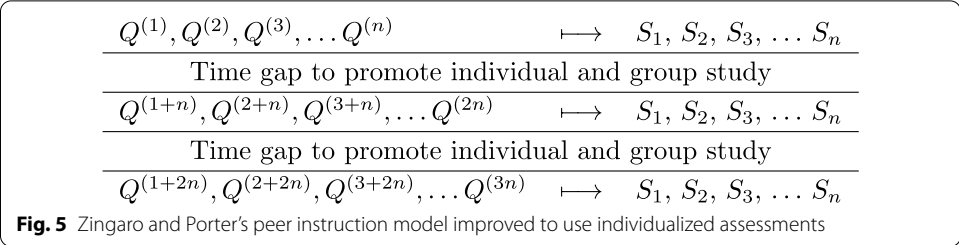
One of the difficulties students may have when they encounter isomorphic questions is that they may not realize that the questions are isomorphic. If the students do not recognize the underlying similarity, the questions are likely to be more difficult to solve. In her two-part paper, Singh compares the performance of the students who chose to do both the  $Q^{(1)}$  set as well as  $Q^{(2)}$  set against those who chose to do one only of the sets (Singh 2008a, b). She also investigates whether the first group of students, who chose to do both sets, understood the underlying similarity of the isomorphic questions. The sets were a mix of quantitative and conceptual questions, and therefore realizing the similarity required insight into the topic.

### Individualized assessment

Individualized assessments are used in different contexts such as adaptive learning (Zare 2011) and plagiarism mitigation (Manoharan 2017). Individualized assessment frameworks typically utilize isomorphic questions.

There are three typical approaches employed by individualized assessment frameworks:

1. Databank – a question is chosen randomly from a bank of (possibly isomorphic) questions. Coderunner (Lobb and Harlow 2016), for example, uses this approach. The main limitations of this approach are the finite number of question variants in the databank and the time it will take to create a large number of questions.
2. Parameter-varying – where relevant parameters, data or other relevant question inputs are randomly generated subject to appropriate constraints. OASIS (Smaill 2005) uses this approach. A parameter-varying approach can potentially yield a very large number of question variants from a single question template. However, it is limited by the inability to express complex constraints or relationships among the parameters. For example, the questions shown in Figs. 6, 7 would be difficult to express using this approach.
3. Macro – where parts of the questions are marked as macros which at generation time are substituted with the result of executing the macros. Dividni (Manoharan 2019) and the *R exams package* (Zeileis et al. 2014) use this approach. The macro approach is more powerful than the databank or parameter-varying approaches since it allows an instructor to set up complex question and answer patterns. The down-



26 A cipher text, thought to be encrypted using a columnar transposition cipher, has been discovered: AGIWEZOIEDEZNSLREDTNBTTZ  
 What is the THIRD word of the corresponding plain text?  
 (A) INSERT  
 (B) SINGLE  
 (C) BEND  
 (D) AD  
 (E) EDITION

27 It is thought that an image file hides a message using LSB steganography. The RGB values of the first few pixels of the image are given below (expressed in hexadecimal). 0x60A18C, 0x6DA780, 0x6AA186, 0x67A47B, 0x5BA883, 0x66A77C, 0x619887, 0x619A7F  
 What is likely to be the first ASCII character of the hidden message, if the message is ASCII-encoded?  
 (A) X  
 (B) U  
 (C) Z  
 (D) C  
 (E) Y

**Fig. 6** Individualized version of two questions from the first test for course 2xx

side, however, is that the instructor should be able to write these macros (i.e., program fragments).

**Methodology**

To ensure student participation in the isomorphic assessment, we split the midterm test, which typically has a weight of 20%, into two isomorphic tests worth 10% each. The repeat test was conducted a week after the first test (Fig. 5).

We conducted these isomorphic midterm tests in two large computer science classes. The first is a computer systems course at year 2 (course 2xx) with a little over 300 students. The second is a software course at year 3 (course 3xx) with almost 400 students. For expediency of marking, both tests were multichoice. Students received their score for the first test within three days, along with (individualized) answers.

The test questions were formulated using an HTML template and a set of macros, and the test scripts were generated using Dividni (Manoharan 2019). The test for course 2xx had 22 questions while course 3xx had 20. The macros for course 2xx had about 2500 lines of code, while the macros for course 3xx had around 2400 lines.

The questions in the repeat test were all parameter-varying isomorphic versions of the questions in the first test. Figs. 6, 7 illustrate two such sets of isomorphic questions from the tests conducted in course 2xx. The repeat tests used the same HTML template and macros used in the first test, but were generated using different random seed values to ensure that students were very unlikely to ever receive the exact same

- 26 A cipher text, thought to be encrypted using a columnar transposition cipher, has been discovered: YRRDIZNAAAHZAUSBYZMGDELT  
What is the THIRD word of the corresponding plain text?  
(A) BEARD  
(B) TRAILER  
(C) ARE  
(D) AGE  
(E) READILY
- 27 It is thought that an image file hides a message using LSB steganography. The RGB values of the first few pixels of the image are given below (expressed in hexadecimal). 0x7AC77E, 0x7CD77C, 0x70C981, 0x76D77F, 0x6DCB73, 0x77D482, 0x7AC683, 0x77D072  
What is likely to be the first ASCII character of the hidden message, if the message is ASCII-encoded?  
(A) J  
(B) E  
(C) A  
(D) M  
(E) I

**Fig. 7** Individualized version of two questions from the repeat test for course 2xx

question twice. While coding the macros took considerable time, generating the test scripts takes only a few minutes.

Since the material in the first half of the course is examined in the tests, the final exam focused more on the second half of the course. Among the small number of exam questions that tested the first half of the course, we included some isomorphic questions from the midterm tests so that we could determine if the students retained the knowledge reinforced through the repeat testing.

We monitored the class forums for the number of discussions related to the test, and student performance data. The performance data consisted of the complete (anonymous) record of each student’s marks for each question in the two tests as well as the marks for those small number of exam questions isomorphic to questions that appeared in the tests. This data would reveal if the students performed better in the repeat test in each of the questions, as well as if students retained their knowledge to perform well in those isomorphic questions in the exam. Furthermore, in course 2xx, 15 of the 22 test questions were related to assignment material that was completed prior to the first test, and it was of interest to determine how this influenced performance in both tests.

In addition, we ran surveys asking students to rate the repeat test. The following questions with answers in five-level Likert scale (strongly disagree to strongly agree) were asked of the students:

1. The repeat test helped me to learn the concepts that I hadn’t adequately learned for the first test.
2. The repeat test helped me to improve my score.
3. It would be a good idea to have a repeat test in other courses.
4. Overall, I like the idea of a repeat test.

The survey also included a free-format response question which asked the students if they had anything else to share.

The class forum usage and overall grade distributions of the two courses were compared with those from the same course in the previous year. The sole structured difference in the teaching methodology was the use of repeat individualized assessment.

**Student satisfaction**

Students’ responses to the questions about the repeat test were largely positive (Tables 1–2), and also show that the majority of students would like to have repeat tests in other courses.

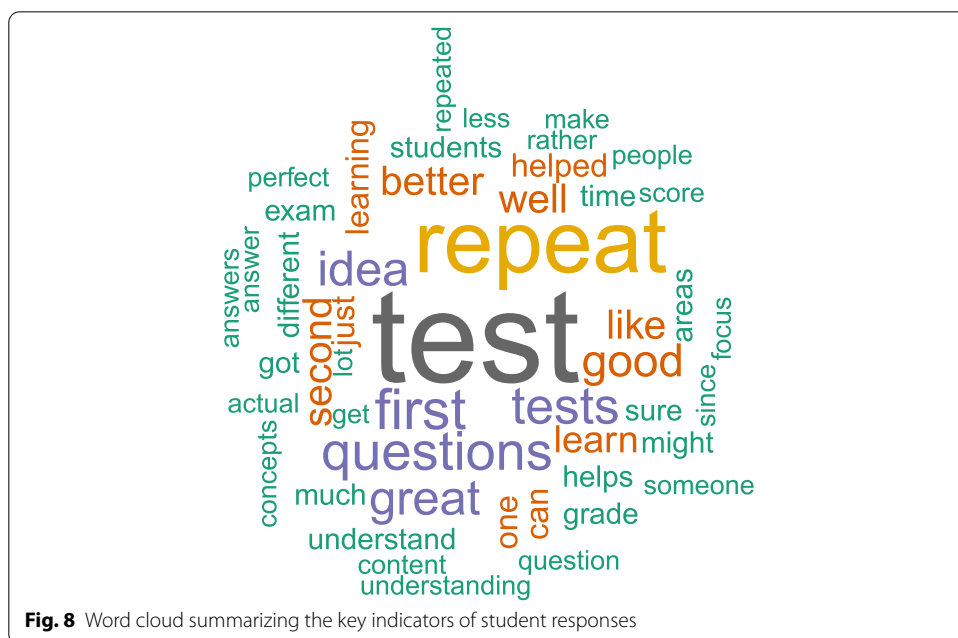
A word-cloud content analysis of the free-format responses was used to summarize the key indicators within the students’ responses (Fig. 8). The most prevalent

**Table 1** Student evaluation results of the repeat test (course 2xx). SD: strongly disagree; D: disagree; N: neutral; A: agree; SA: strongly agree

Question	SD (%)	D (%)	N (%)	A (%)	SA (%)
The repeat test helped me to learn the concepts that I hadn’t adequately learned for the first test	5.8	11.5	6.9	26.4	49.4
The repeat test helped me to improve my score	3.5	2.3	6.9	12.6	74.7
It would be a good idea to have a repeat test in other courses	13.8	4.6	23.0	13.8	44.8
Overall, I like the idea of a repeat test	13.8	11.5	12.6	16.1	46.0

**Table 2** Student evaluation results of the repeat test (course 3xx). SD: strongly disagree; D: disagree; N: neutral; A: agree; SA: strongly agree

Question	SD (%)	D (%)	N (%)	A (%)	SA (%)
The repeat test helped me to learn the concepts that I hadn’t adequately learned for the first test	2.6	3.5	8.2	16.5	69.4
The repeat test helped me to improve my score	1.2	4.7	7.1	12.9	74.1
It would be a good idea to have a repeat test in other courses	4.7	3.5	10.6	20.0	61.2
Overall, I like the idea of a repeat test	5.9	3.5	8.2	21.2	61.2



**Fig. 8** Word cloud summarizing the key indicators of student responses

adjectives were “great”, “good” and “better”, indicating that they were overwhelmingly positive about the repeat test. Some responses explicitly confirmed that it helped them to improve their learning and understanding of concepts. Examples include:

- “Repeat tests seemed like a good way to encourage iterative revision. It exposed areas where I hadn’t prepared as well as I thought I had, and gave me a chance to focus on those areas in preparation for the second test.”
- “This was also a great idea because university is all about learning and since it’s just a test and not worth a lot of our grade like the final exam it’s okay that generally students will get a higher grade in the second test. I definitely found that it was great because I could work on the areas I got wrong. I personally made a number of dumb mistakes on easy questions in my first test which didn’t reflect my actual knowledge. The second test can help to represent an actual representation of my knowledge by averaging out the grades of the two tests.”
- “I like how I got almost perfect score. Sometimes tests come from wide frame of coursework and something I might not remember or even understand very well might come out. With repeated test I have fair chance of showing my learning progress.”
- “It was good and helped me review important things I missed the first time.”
- “I thoroughly agree with repeat tests over just one single test because for someone like me, I’m not a good test taker, so I generally don’t do well in those environments. So, to repeat the test helped my grade hugely.”
- “The repeated test was great, gave me the confidence to learn and I actually know the content better, as do a lot of people in class, a repeated test pretty much helps you not freak out about a theory mark and helps you focus content and actually learn something, thank you.”
- “Great idea. Really helped me focus on the important course material.”
- “Good idea as helps secure the concepts needed.”
- “Great idea, I learned a lot from it!”

Other keywords in the word-cloud were also investigated for their usage within the free-format responses. One of particular interest was “different”, which mainly was used in responses pertaining to the degree of difference between the two tests. Examples include:

- “It would be good if the repeat test were a bit more different than the first one, or at least introduce some more new questions that were not like the ones in the first one.”
- “I like the idea of retesting the same content, but if the questions were basically identical, meaning that students only needed to go learn how to answer those questions, rather than go learn the actual concepts and gain an understanding of the questions – basically they could rote learn them for the second test, even with different variables.
- “Vary the questions slightly more in the second test so it is necessary to learn the concepts behind the question rather than just the question itself.”



The issue of providing adequate variation between the two tests is discussed in Section 6.1.

**Analyses**

**Peer discussion on online class forums**

There were many more post-test discussion threads after the (first) test when there were repeat tests, confirming the possibility of enhanced peer learning at play (Table 3). Pearson  $\chi^2$  tests established that these increases were both highly statistically significant ( $p$ -value < 0.001).

**Statistical methodology to analyse test and exam performance**

The binary outcomes (correct or incorrect) of the questions on the two midterm tests and final exam are not statistically independent due to multiple outcomes measured on each student and the repeated use of variants of each isomorphic question. Hence, a repeated measures form of contingency table analysis was performed by using the `glmmTMB` function within the R language to fit mixed-effects logistic regression models. These models included student and question as random effects.

It was felt that overall student ability may have an effect on the learning pathway, and so as a measure of overall ability the students in each class were evenly split into two ability groups depending on whether their final exam score was above (high ability) or below (low ability) the median final exam score.

The following questions were examined:

- What effect does overall student ability have on the probability of answering a question correctly in Test 1? Does having previously seen the question in an assignment (course 2xx only) also have an effect?
- Was there an improvement from Test 1 to Test 2? Moreover, does this effect depend on overall student ability, or whether the question was previously seen in an assignment (course 2xx only).
- How did test performance influence final exam performance?

In what follows, the notation Ability, Asgmt, T1, and T2 is used to denote the variables corresponding to student ability (high or low), whether the question material was seen in an assignment prior to Test 1 (yes or no), Test 1 question outcome (correct or incorrect), and Test 2 question outcome (correct or incorrect).

**Table 3** Class forum discussion threads about tests in the two courses after the (first) test – 2017 had no repeat tests while 2018 did

Course 2xx		Course 3xx	
2017	2018	2017	2018
2	26	1	23

**Course 2xx results**

***Analysis of test 1 performance***

Analysis of Test 1 results included the explanatory variables Ability and Asgmt. Over all students and questions, the proportion of Test 1 questions answered correctly was 0.520. Whether or not the question material was previously-seen in assignments made little difference, with success rates of 0.529 if previously seen, and 0.501 if not. This difference was not statistically significant ( $p > 0.05$ ). For students of high ability, the log-odds of answering a question correctly was 0.743 higher than for low ability students ( $p < 0.001$ ).

***Analysis of test 2 performance***

Comparison with Test 1: The overall success rates on the questions in Tests 1 and 2 were 0.520 and 0.775 respectively, and the improvement was highly significant ( $p < 0.001$ ).

Additional effects of ability and assignment: The full analysis of Test 2 results included the explanatory variables Ability, Asgmt and T1. Students with high ability had log-odds for a correct answer that was 0.643 higher than students of low ability ( $p < 0.001$ ), and this effect of Ability was independent of Asgmt and T1 (Table 4). The variables T1 and Asgmt were both highly significant, as was their interaction ( $p < 0.01$ ).

Answering a T1 question correctly increased the odds of answering the same T2 question correctly, and having previously seen the question material in an assignment also increased the odds. Furthermore, the interaction between Asgmt and T1 showed that there was an additional positive benefit from having both answered the question correctly in T1 and having previously seen the question material in an assignment (Table 4).

***Analysis of exam performance***

Two test questions were repeated in the exam, both on material previously seen in assignments prior to Test 1. These two questions had the highest score over all of the exam questions, with combined success rate of 0.941. Due to the limited data, the analyses were numerically unstable if multiple explanatory variables were used and so only T2 was used, and was statistically significant ( $p < 0.001$ ). Combined over the two questions, students who answered an isomorphic T2 question incorrectly had an 0.865 success rate, increasing to 0.964 for those who answered correctly in T2. This corresponds to an increase in log-odds of 1.54.

**Table 4** Effect of Ability, Asgmt and T1 on the log-odds of correctly answering an isomorphic Test 2 question. The baseline is for a low-ability student who incorrectly answered the question in Test 1, for a question not previously seen in an assignment

Effect	Estimate	Std error
Baseline	0.467	0.256
Ability=high	0.643	0.143
T1=correct	0.389	0.119
Asgmt=yes	0.809	0.292
(T1=correct)*(Asgmt=yes)	0.407	0.156

**Course 3xx results**

***Analysis of test 1 performance***

None of the questions in the tests were previously seen in assignment material, so Ability was the only explanatory variable. High ability students were found to have 0.678 higher log-odds of answering a Test 1 question correctly ( $p < 0.001$ ) compared to low ability students.

***Analysis of test 2 performance***

Comparison with Test 1: The overall success rates on the questions in Tests 1 and 2 were 0.665 and 0.879 respectively, and the difference was highly significant ( $p < 0.001$ ).

Additional effect of ability:

The full analysis of Test 2 results included explanatory variables Ability and T1. Students with high ability had 0.741 higher log-odds of answering correctly ( $p < 0.001$ ) compared to low ability students, and answering the Test 1 question correctly increased the log-odds of answering the Test 2 question correctly by 0.475 ( $p < 0.001$ ). There was no interaction between Ability and T1 (Table 5).

***Analysis of exam performance***

Three test questions were repeated in the exam, and performance on these three questions was significantly better than on the other questions ( $p < 0.001$ ), with a combined success rate of 0.932. As with the course 2xx analysis, it was only possible to use T2 as an explanatory variable, and it was statistically significant ( $p < 0.001$ ). Combined over the three questions, students who answered an isomorphic T2 question incorrectly had a 0.773 success rate, increasing to 0.946 for those who answered correctly in T2. This corresponds to an increase in log-odds of 1.63.

**Comparison of grade distributions**

Table 6 shows the raw (unscaled) grade distributions for the two courses in years 2017, where there was no repeat assessment, and 2018. The 2017 grade distributions for both courses were skewed towards lower grades, especially the 2xx course with a grade average below 3 (C+). Consequently, in 2017, the raw grades for both courses shown in Table 6 were manually scaled up to improve the distribution. The 2018 grade distributions were seen to be more bell-shaped, and with improvements in grade average of 0.92 and 0.42 in 2xx and 3xx, respectively. No adjustments were made to the 2018 raw grades.

For each course, a Pearson’s  $\chi^2$  test was performed to determine whether there was a significant difference between the 2017 and 2018 grade distributions. Pearson’s  $\chi^2$  tests

**Table 5** Effect of Ability and T1 on the log-odds of correctly answering an isomorphic Test 2 question

Effect	Estimate	Std error
Baseline	2.377	0.382
Ability=high	0.741	0.125
T1=correct	0.475	0.092

The baseline is for a low-ability student who incorrectly answered the question in Test 1

**Table 6** Raw grade distributions in the two courses – 2017 had no repeat assessments while 2018 did

Grade	Grade Point	Course 2xx		Course 3xx	
		2017	2018	2017	2018
A+	9	2	16	8	11
A	8	5	22	24	19
A-	7	11	22	39	45
B+	6	31	25	43	47
B	5	47	42	67	67
B-	4	63	43	65	66
C+	3	81	39	48	55
C	2	65	29	29	43
C-	1	40	16	20	16
Fail	0	71	56	74	31
Total Grade		416	310	417	400
Average		2.94	3.86	3.87	4.29

13 Consider the following JavaScript code block that uses destructuring to extract elements of a list.

```

const fn = (list) => {
  const [h, ...t] = list;
  if (t !== undefined && t.length !== 0)
    const tmp = fn(t);
    return h < tmp ? h : tmp;

  else return h;
};
    
```

What is the value of `fn([ 6, 11, 14, 14, 12, 7, 12, 1])`?

(A) 12  
 (B) 7  
 (C) 6  
 (D) 5  
 (E) 14

**Fig. 9** A question that is prone to rote learning – Course 3xx

yielded a *p*-value of < 0.001 for course 2xx and a *p*-value of 0.007 for course 3xx, both providing highly significant evidence of a difference.

**Limitations**

**Rote learning**

While the generation of the repeat test is essentially free, if the questions aren’t designed well with the repeat test in mind then the students could rote learn the process of just deriving the answer without understanding the concept. By way of example, Figure 9 shows a question used in the first test of course 3xx. The learning outcome in this case is to understand the *destructuring* feature of lists in the context of a given function, and then to work out the answer when applying this function to some supplied data (that varies between versions). The function in this particular case finds the maximum value in a given list of integers. In a repeat test, however, if we used the same function, a student could simply pick up the maximum value without having to understand how the

*destructuring* feature works in the context of the function. The repeat test therefore implemented a different binary comparison operator (on code line 5) to ensure that the students need to show that they were able to meet the learning outcome rather than rote-learn.

If there is a possibility that the students may rote learn the answer then the isomorphic questions need to have more variability than provided by just changing a simple parameter such as the data. In the above example of the destructuring question (Fig. 9), we regard the binary comparison operator as an additional input that can be varied between parameter-varying isomorphic versions of this question. In this case, it would vary between the first test and repeat test, but not at an individualized level, and hence would not be a variation anticipated by students studying for the repeat test. We suggest such use of multi parameter-varying isomorphic questions as a useful tool when rote learning is a concern. More generally, wider classes of isomorphic question could be used, such as different versions of the wording or presentation of the question.

### **Fairness**

It is quite important to ensure that the parameter-varying isomorphic questions have an equal level of difficulty. Otherwise it will be unfair. For example, while the questions “what is the sum of 2 and 8?” and “what is the sum of 956 and 775?” are isomorphic, one can see that the latter variant is harder and more time-consuming. This issue can be answered by limiting the range of the parameters. In the example here, we can choose a single digit number between 2 and 9 inclusive; we exclude 0 and 1 because they will be easier than other single-digit choices.

### **Conclusions**

We have used a macro-based individualized assessment framework to generate parameter-varying isomorphic tests that proved to be pedagogically successful in two large classes. While it took time to set up questions for the first test, the instructors required very little preparation time for the repeat test.

This application of repeat testing did not include any formal peer instruction, but it was observed that activity on the class forum was greatly increased after the first test compared to the previous year in which there was no repeat test. Thus, the total beneficial effect of repeat testing also includes the effect of any increased online peer-assisted learning (Watts et al. 2015) focused around the questions in the tests.

In the main, students rated their experience with repeat testing very highly, with about 75% neutral or favourable about the idea of repeat tests and more than 80% either neutral or favourable about their use in other courses. While many responses reported that it was a “great” or “good” idea that aided their learning, a minority indicated that they would like to see slightly greater variability in the isomorphic questions.

The statistical analyses confirmed that the repeat isomorphic assessments enabled positive learning. Not surprisingly, in both classes, students with higher ability (as measured by exam success) performed better on Test 1 and Test 2. The analyses of Test 2 results provided the interesting result that the relative effect of answering a Test 1

question correctly did not depend on overall ability. This learning was retained in the exam, with higher marks being scored on the previously seen test questions.

The data from course 2xx provided interesting insight into the effect of having previously seen the test question material in an assignment. There was no significant effect on Test 1, but it did have a very strong effect on the way that students learned between taking Test 1 and Test 2. During that focused period of study, previously seeing material in an assignment greatly increased the odds of correctly answering in Test 2. This unanticipated effect may be due to the online peer discussion being largely unsupervised, and hence the assignments would be of great help to learn the concepts. This was echoed in the student evaluations – where students felt the repeat Test 2 helped them to focus on concepts that they had not adequately learned for Test 1. The benefit was stonger for students who had answered the question correctly in Test 1.

Our quantitative statistical analyses quantified the strength of the positive learning. For example, a correct Test 2 answer increased the log-odds of a successful exam answer by 1.54 in course 2xx, and 1.63 in course 3xx. These values are not significantly different, demonstrating consistency in the positive learning across the two classes. Moreover, our methodology could be applied in further studies to find improvements for increasing the strength of positive learning using repeat assessments, such as the inclusion of formal peer instruction.

#### **Acknowledgements**

Not applicable.

#### **Funding**

Not applicable.

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

Not applicable.

#### **Declarations**

#### **Ethics approval and consent to participate**

Not applicable.

Received: 22 October 2020 Accepted: 15 March 2021

Published online: 27 April 2021

#### **References**

- Singh, C. (2008a). Assessing student expertise in introductory physics with isomorphic problems. I. Performance on non-intuitive problem pair from introductory physics. *Physical Review Special Topics-Physics Education* 4 010104. <https://doi.org/10.1103/PhysRevSTPER.4.010104>.
- Singh, C. (2008b). Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer. *Physical Review Special Topics-Physics Education* 4 010105. <https://doi.org/10.1103/PhysRevSTPER.4.010105>.
- Zingaro, D., Porter, L. (2015). Tracking student learning from class to exam using isomorphic questions. In: Proceedings of the 46<sup>th</sup> ACM Technical Symposium on Computer Science Education. SIGCSE '15, pp. 356–361. ACM, New York, NY, USA. <https://doi.org/10.1145/2676723.2677239>
- Kjolsing, E., & Einde, L. V. D. (2016). Peer instruction: Using isomorphic questions to document learning gains in a small statics class. *Journal of Professional Issues in Engineering Education and Practice*, 142(4), 04016005. [https://doi.org/10.1061/\(ASCE\)EI.1943-5541.0000283](https://doi.org/10.1061/(ASCE)EI.1943-5541.0000283).
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970–977. <https://doi.org/10.1119/1.1374249>.
- Lobb, R., & Harlow, J. (2016). Coderunner: A tool for assessing computer programming skills. *ACM Inroads*, 7(1), 47–51. <https://doi.org/10.1145/2810041>.
- Manoharan, S. (2019). Cheat-resistant multiple-choice examinations using personalization. *Computers & Education*, 130, 139–151. <https://doi.org/10.1016/j.compedu.2018.11.007>.
- Small, C. R. (2005). The implementation and evaluation of OASIS: A web-based learning and assessment tool for large classes. *IEEE Transactions on Education*, 48(4), 658–663. <https://doi.org/10.1109/TE.2005.852590>.

- Zingaro, D., Porter, L. (2014). Peer instruction in computing: The value of instructor intervention. *Computers & Education* 71, 87–96.
- Zare, S. (2011). Personalization in mobile learning for people with special needs. In: Stephanidis, C. (ed.) *Universal Access in Human-Computer Interaction. Applications and Services. Lecture Notes in Computer Science*, vol. 6768, pp. 662–669. Springer, New York. <https://doi.org/10.1007/978-3-642-21657-2-71>.
- Manoharan, S. (2017). Personalized assessment as a means to mitigate plagiarism. *IEEE Transactions on Education*, 60(2), 112–119. <https://doi.org/10.1109/TE.2016.2604210>.
- Zeileis, A., Umlauf, N., & Leisch, F. (2014). Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond. *Journal of Statistical Software, Articles*, 58(1), 1–36. <https://doi.org/10.18637/jss.v058.i01>.
- Watts, H., Malliris, M., & Billingham, O. (2015). Online peer assisted learning: Reporting on practice. *Journal of Peer Learning*, 8, 85–104.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---