ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE
Paper ID #37256

Is performance on tests affected by the difficulty of the first question and an informational message about the benefits of the 'testing effect'?

Bruno Korst (Assistant Professor, Teaching Stream)

Bruno Korst completed his undergraduate studies at the Faculdade de Engenharia Industrial (FEI), in Brazil, where he received the title of Electrical Engineer. In Canada, he completed his Master's at Carleton University, specializing in signal processing for acoustics, and worked in the industry in this field. He started his work at the University of Toronto as a member of technical staff and served as the Director of Teaching Laboratories for ECE prior to joining the Faculty as a Teaching Stream professor. He has received multiple awards on innovation, and was the first staff member to receive the Gordon R. Slemon Award for excellence in the teaching of design. Motivated by his strong interest in laboratory teaching within engineering education, he is presently completing a PhD in Cognitive Neuroscience at the University of Waterloo, with his research concentrating on prospective attention as applied to video instruction. In addition to his technical training and practice, he also holds a B.A. in Political Science/Int'l Relations (Calgary) and an MBA in Marketing (FGV – Brazil). He is a licensed Professional Engineer in the province of Ontario.

Dan Wolczuk

Dan Wolczuk is a Lecturer in the Faculty of Mathematics at the University of Waterloo. He primarily teaches calculus and linear algebra and conducts research in the scholarship of teaching and learning. In 2022 he won the Canadian Mathematical Society's Excellence in Teaching Award and in 2021 he won the University of Waterloo's Distinguished Teacher Award. His website is wolczuk.com.

Daniel Smilek (Professor)

Is performance on tests affected by the difficulty of the first question and an informational message about the benefits of the 'testing effect'?

Bruno Korst[1,2], Dan Wolczuk[3], Dan Smilek[1]
1- Dept. of Psychology, University of Waterloo
2- Dept. of Electrical Engineering, University of Toronto
3 – Faculty of Mathematics, University of Waterloo

Abstract

This paper presents evidence-based practice applied to course design and delivery, through a study conducted during an in-person undergrudate course exploring several aspects of test delivery. An undergraduate linear algebra course was initially designed to draw on the benefits of the well-documented testing effect, which is characterized by better student learning as a result of frequent testing. A study was conducted over one semester with the goal of assessing objectively whether the addition of a lecture-long informational message about the testing effect delivered by the instructor could enhance overall performance. In addition, the study aimed to investigate an aspect of test design, namely whether the difficulty of the first question (easy vs. hard) would affect the overall performance on the tests.

The cohort consisted of 119 students of different STEM areas across a number of sections, all taught by the same instructor. The course included a total of 9 quizzes, 8 of which relevant to the study, each consisting of 3 questions that varied in difficulty; four quizzes started with a hard question and four started with an easy question. The course also included a midterm test and a final exam. The cohort was divided into two counterbalanced groups with one counterbalance receiving the easy question first on odd numbered quizzes and hard questions first on the even numbered quizzes, and with the other counterbalance group experiencing the reverse. All quizzes and exams were delivered at appropriately scheduled times to all students and the same amount of time was given to all students to solve the questions on the quizzes. Critically, one section of the course was chosen to receive an informational message about the testing effect explaining how frequent testing improves performance and encouraging students to use the quizzes as a learning opportunity. For this one section, the informational message was delivered once, after the first quiz (second week in the term). All students received messages of encouragement from the instructor throughout the term.

Results showed significantly higher performance on the easy questions than the hard questions indicating the manipulation of question difficulty was successful. However, there was no difference in performance between those participants for whom the quiz started with an easy question than those for whom the quizzes started with a hard question. Notably, grades were higher for the group that received the motivation message than the group that did not receive the message. It is hoped that this promising result can be extended in future experiments, which may include multiple informational messages about the effectiveness of testing throughout the term.

Introduction

Large courses with multiple sections and instructors are common in the first few years of many university undergraduate programmes. In such courses, assessing student learning can be challenging. For many such courses, assessment primarily takes the form of mid-term tests and a final exam, with the possibility of quizzes throughout the term. When in person, such tests are often taken by students at the same time, in a large assembly within a predetermined time. While assessments of learning are necessary in courses to guage student learning, they may take away from class instruction time. Thus, it is imperative that the processess of testing be well understood so that negative aspects can be reduced and the postive aspects leveraged. Here we focus on several issues related to assessing student learning during such large undergraduate courses.

Although testing may take away from class learning time, studies have shown that frequent testing is beneficial to student learning [1][2]. The benefits stem both directly from the test itself as the student retrieves information during the multiple tests (and that enhances retention [3]) as well as indirectly, as the student tested frequently is more prone to engage in regular studying [4]. The benefit that testing has on learning is referred to as the *testing effect* [1]. Tests also inform students what they know and what they do not know, therefore improving their ability to predict their future performance and concentrate on studying the material which they do not know so well. One issue that has not been fully addressed is whether providing students with information about the testing effect has the potential of increasing the benefits of testing, perhaps by motivating students to construe and focus on tests as an important learning opportunity. Anecdotally, the experience of one instructor of a second-year math course (one of the authors) suggests that on previous iterations of the course, learning was better when he first informed students about the benefits of testing than when he did not.

In addition to the frequency of testing, the ordering of test questions in terms of their difficulty might also influence performance. Some have advanced the notion that tests with questions ordered in increasing levels of difficulty lead students to perform better than tests with other orders of questions in terms of their difficulty. This view may have been based on early adopted practices [5]. While there does seem to be some evidence consistent with this notion [11][12][13], the evidence is weak (as in [12]). Furthermore, a number of studies have investigated the order of question difficulty as well as order of topics relatively to random tests and have found no significant effects on test performance [5]-[10]. Given that these studies often made use of long tests, drawing tests from established question banks for their respective areas, it remains to be seen whether short quizzes starting with and easy question or with a hard question will result in a significant influence on student performance.

The Present Study

The present study extended prior work in two ways: First, we assessed whether the cognitive benefits of testing can be increased by presenting students with a informational message about

the benefits of frequent testing and the importance of using quizzes as a learning opporunity. Second, building on prior work examining the impact of question difficulty, we evaluated whether the difficulty of the first question on quizzes (i.e., being either hard or easy) impacts students performance on the quizzes. Based on prior work, we expected either no effect of first question difficulty, or possibly that students will show better overall performance when quizzes start with an easy question than when they start with a harder question.

Experiment

The experiment involved 119 consenting participants who were all students regularly enrolled in a second-year linear algebra course at the University of Waterloo in Ontario, Canada. The study involved five sections of the course, all taught by the same instructor. The course was designed to have a total of nine quizzes through the term, as well as a midterm and a final exam. Each quiz had three questions of different difficulty levels. The quizzes, the midterm test and the final exam were worth, respectively, 12%, 28% and 60% of the final grade. For the analyses below, only responses and grades from consenting students were used.

Importantly, we manipulated the presentation of information about the testing effect across the five lecture sections of the course. One of the five lecture sections of the course (the "Informational message" (IM) group) was chosen to receive a message informing students of the proven benefits of the testing effect. This group was comprised of 40 consenting participants. The informational message about the testing effect was delivered once, by the instructor, during the lecture prior to the second quiz (close to the beginning of the term). The specific message about the testing effect was thorough, covered a significant part of the lecture for the day, and included academic references to research showing the benefits to the students of having multiple tests/quizzes throughout the academic term. The remaining four lecture sections did not receive this information about the benefits of testing (the "No informational message" (NIM) group). More specifically, as shown in Figure 1, Section 5 of the course received information about the testing effect whereas Sections 1 to 4 did not.

All students received deliberate but general, brief, encouraging comments throughout the semester to continue with their efforts to keep up with the lecture material presented. These encouragement messages remarked on the results of the previous quiz and encouraged students to keep up with their preparation for the next quiz. There were no pre-defined scripts to the messages, and it was left up to the instructor how to better deliver it during the lectures.

The ordering of the questions with regard to question difficulty was varied across quizzes. To counterbalance the question difficulty order on the quizzes, participants in each of the five sections of the course were split into two counterbalanced groups. The first quiz included the same question order for both groups. As shown in Figure 1, after quiz 1 (Q1), the two counterbalanced groups – C1 and C2 – were given different alternating sequences of quizzes. For group C1, quizzes Q2, Q4, Q6 and Q8 started with an easy question, whereas the other quizzes started with a hard question. Group C2 was given a sequence of quizzes with an order of first question difficulty opposite to that of group C1. Thus, following the first quiz, both groups

completed a total of eight of quizzes alternating in terms of first question difficulty, four starting with a hard question and four with an easy question. Figure 1 also shows the placement of the midterm test (MT) and the final exam (FE).

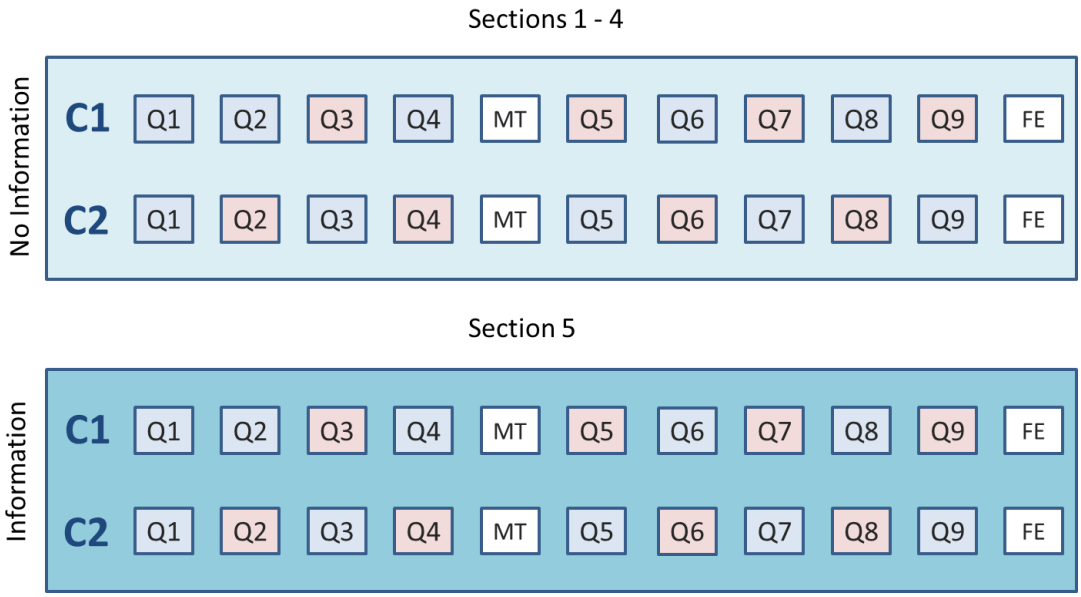


Figure 1 – Experimental layout: Sections 1-4 (did not receive information about the testing effect— NIM Group), Section 5 (received information about the testing effect— IM Group), Counterbalances C1 and C2, Quizzes Q1-Q9, Midterm test (MT) and Final exam (FE).

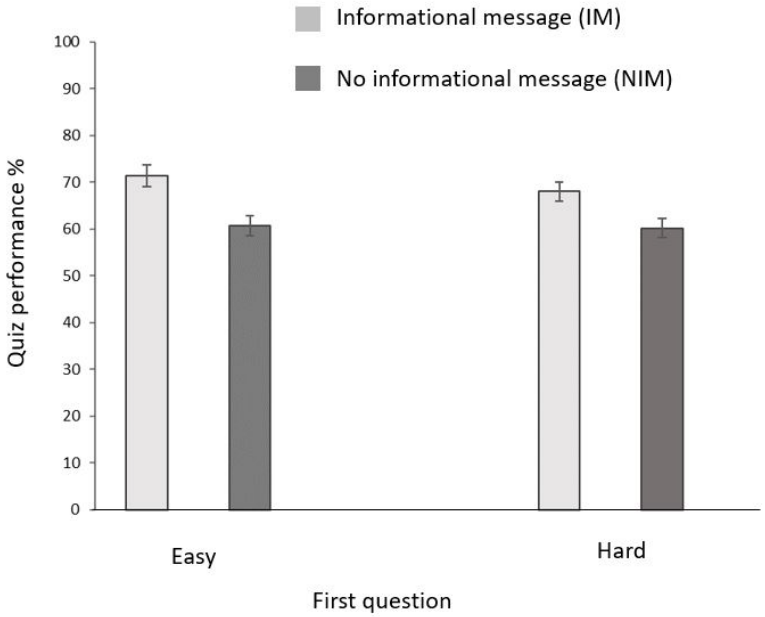
It should be further noted that the same instructor who delivered the messages and lectures, also designed the quizzes and tests. All quizzes and tests were delivered at the same time to all students and all students were given the same amount of time to solve the quizzes and tests. Furthermore, the time constraint was such that it did not allow for much free time. Finally, seats were assigned per student number such that the group which received information on the testing effect and the group which did not were mixed in assigned seats, with no distinction between the groups during the time of the quizzes and tests.

Results and Discussion

As a manipulation check, we examined scores across quiz questions deemed to be easy and those deemed to be hard. A repeated-measures t-test revealed that the performance was higher for the easy questions ($M = 73.5\%$, $SE=1.54$) than for the hard questions ($M =55.9\%$, $SE=1.73$), $t(118)=16.02$ $p < 0.001$.

We next analyzed average quiz performance of the eight quizzes following the first first quiz as a function of Informational Message and the Difficulty Order of the questions (see Figure 2). The average quiz scores for each participant in each condition were entered into an Analysis of Variance (ANOVA) with Informational Message (IM vs. NIM) as a between-participant factor and question Difficulty Order (Easy First, vs. Hard First) as a within-participant factor. The

analysis revealed a significant main effect of Informational Message ($F(1,117) = 9.58$, $p<0.001$), such that quiz scores were higher for those who received information about the testing effect than those who did not. Neither the main effect of Difficulty Order ($F(1, 117) =1.012$, $p=0.32$) or the interaction ($F(1, 117)=.87$, $p=0.35$) reached statistical significance.



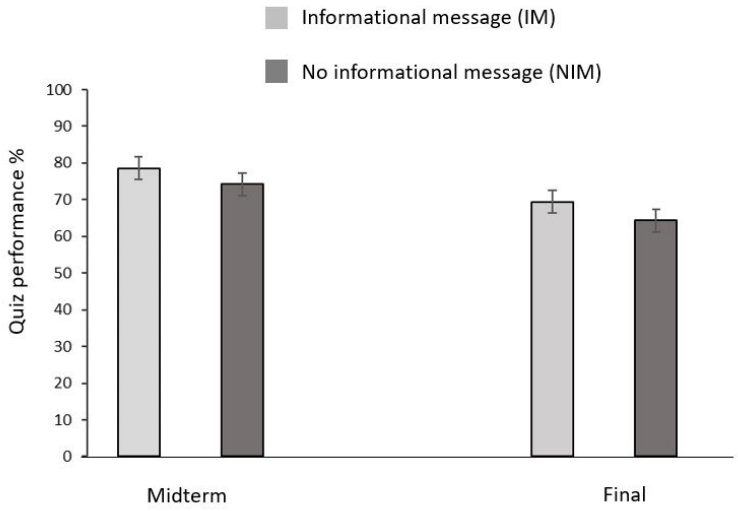
	Mean	Std Error
Easy - IM	71.44	2.34
Hard – IM	67.99	2.03
Easy - NIM	60.69	2.20
Hard - NIM	60.17	2.08

Figure 2. Mean quiz performance (%) as a function of having received the Informational Message (IM vs. NIM) and the Difficulty Order (First Question Easy vs. Hard) of the first question. The error bars reflect one standard error of the mean.

We also examined whether the inclusion of the informational message about the testing effect influenced performance on the midterm and the final exam. The mean scores on the midterm for each Informational Message group and the scores on the final for each of the two groups are shown in Figure 3. The scores were entered into an Analysis of Variance (ANOVA) with Informational Message (IM vs. NIM) as a between-participant factor and Test (midterm test and final exam) as a within-participant factor. The analysis revealed a significant main effect of Test ($F(1,117)=88.21$, $p<0.001$), such that the midterm test marks were higher than those of the final exam. There was a marginal but non-significant effect of Informational Message, $F(1,117)=3.58$, $p=0.061$. There was no effect for interaction, $F(1,117)=0.12$, $p=0.7$.

The scores on the two assessments were further submitted to separate independent sample t-tests with Information Message as the between-participant factor. Since the sample sizes are unequal between the IM and NIM groups, we opted to use Welsh's t-test for correction. The t-tests also revealed a marginal but non-significant effect of Informational Message on the midterm test ($t(83.87)=1.88$, $p=0.06$) and on the final exam ($t(105.1)=1.88$, $p=0.06$). Inspection of Figure 3

shows that the group which received the information once again performed nominally better than the group which did not receive the information.



	Mean	Std Error
Midterm - IM	78.59	1.86
Final – IM	69.46	1.92
Midterm - NIM	74.18	1.43
Final - NIM	64.29	1.95

Figure 3. Mean performance (%) on the Midterm and the Final Test as a function of Informational Message (IM vs. NIM). Error bars reflect one standard error of the mean.

Conclusions

In the present study we examined two questions related to testing in large undergraduate courses. The first question was whether quiz and test performance would be better if students were provided with an informational message about the effectiveness of testing as a tool for learning than if they were not. In this regard, we did indeed find that students' quiz performance was better for those who received information about the testing effect at the beginning of the term. Interestingly, this benefit seemed to also bleed over to the main tests of the course, suggesting that the infromational message had an impact on overall learning. While we cannot fully rule out the possiblity that the section that received the informational message was not composed of better performers even without the message, we think this possibility is unlikely. Rather, we suspect that those who were carefully informed about the benetifs of testing were more diligent than others when preparing for the quizzes (even though they were each of worth only a small fraction of the overall course grade) and were perhaps more motivated to learn from the quizzes than others. It is also worth noting that in this experiment, the information about the testing effect was presented only once at the beginning of the term. It remains to be determined whether a similar – or greater – benefit can be achieved with the information given multiple times through the term. Furthermore, we observed a strong trend in performance improvement on both the midterm test and the final exam for the group which received the informational message. This could be further investigated with the use of multiple informational messages through the term rather than one at the beginning of the term.

The second question addressed by our investigation was whether starting a quiz with an easy or hard question would differentially influence performance on the quizzes. Results presented above indicate that the difficulty of the first question did not significantly affect student performance on the quizzes. These results are inconsistent with anecdotal reports that students perform better on exams that start with easy questions than with hard ones, but they are consistent with other work showing no effect of question diffculty order. It is worth noting that during the quizzes, students were free to decide which question they would attempt to answer first. Some students might have used a personal strategy of trying to identify the easier (or the hard) question to start. During the administration of the quizzes it was informally observed that students tended to solve the questions from beginning to end, without applying a strategy based on difficulty. This sequential approach was anecdotally reported also in previous iterations of the course, and it is plausible that students just answer the questions sequentially given the time constraints of the quiz. However, future research should examine the influence of question difficulty order while holding the sequence of question completion rigidly constant.

References

[1] H. L. Roediger III and J. D. Karpicke, *"Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention"* Psychological Science, 2006 17:249

[2] H. L. Roediger III, A. L. Putnam, M. A. Smith, *"Ten Benefits of Testing and Their Applications to Educational Practice"* Psychology of Learning and Motivation, Vol 55, Chapter One, 2011

[3] M. Carrier, H. Pashler, *"The Influence of Retrieval on Retention"* Memory and Cognition, 20, 633-642, 1992

[4] V. T. Mawhinney, D. E. Bostow, D. R. Laws, G. J. Blumenfeld, B. L. Hopkins, *"A comparison of students studying-behaviour produced by daily, weekly and three week testing schedules"* Journal of Applied Behaviour Analysis, 4, 257-264, 1971

[5] M. H. Brenner, *"Test Difficulty, Reliability and Discrimination as Functions of Item Difficulty",* Journal of Applied Psychology, Vol 48, No. 2, 98-100, 1964

[6] B. S. Plake, *"Item Arrangement and Knowledge of Arrangement on Test Scores"* The Journal of Experimental Education, 49:1, 56-58, 1980

[7] D. L. Newman, D. K. Kundert, D. S. Lane Jr., K.S. Bull, *"Effect of Varying Item Order on Multiple-Choice Test Scores: Importance of Statistical and Cognitive Difficulty"* Applied Measurement in Education, 1:1, 89-97, 1988

[8] D. L Neely, F. J. Springston, S. J. H. McCann, *"Does Item Order Affect Performance on Multiple-Choice Exams?"* Teaching of Psychology, Vol. 21, No. 1, 1994

[9] A.H. Perlini, D. L. Lind, B. D. Zumbo, *"Context Effects on Examinations: The Effects of Time, Item Order and Item Difficulty"* Canadian Psychology, 39:4, 1988

[10] B. A. V. Schee, *"Test Item Order, Level of Difficulty, and Student Performance in Marketing Education"* Journal of Education for Business, 88:1, 36-42, 2013

[11] R. K. Hambleton, R. E. Traub, *"The Effects of Item Order on Test Performance and Stress"* The Journal of Experimental Education, 43:1, 40-46, 1974

[12] W. R. Balch, *"Item Order Affects Performance on Multiple-Choice Exams"* Teaching of Psychology, Vol. 16, No.2, 1989

[13] H. Chen, *"The Moderating Effects of Item Order Arranged by Difficulty on the Relationship between Test Anxiety and Test Performance"* Creative Education, Vol. 3, 328-333, 2012