The Canadian Journal for the Scholarship of Teaching and Learning

Volume 8 | Issue 1 Article 11

March 2017

Does Correct Answer Distribution Influence Student Choices When Writing Multiple Choice Examinations?

Jacqueline A. Carnegie
University of Ottawa Faculty of Medicine, jcarnegi@uottawa.ca

https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

Does Correct Answer Distribution Influence Student Choices When Writing Multiple Choice Examinations?

Abstract
Summative evaluation for large classes of first- and second-year undergraduate courses often involves the use of multiple choice question (MCQ) exams in order to provide timely feedback. Several versions of those exams are often prepared via computer-based question scrambling in an effort to deter cheating. An important parameter to consider when preparing multiple exam versions is that they must be equivalent in their assessment of student knowledge. This project investigated a possible influence of correct answer organization on student answer selection when writing multiple versions of MCQ exams. The specific question asked was whether the existence of a series of four to five consecutive MCQs in which the same letter represented the correct answer had a detrimental influence on a student's ability to continue to select the correct answer as he/she moved through that series. Student outcomes from such exams were compared with results from exams with identical questions but which did not contain such series. These findings were supplemented by student survey data in which students self-assessed the extent to which they paid attention to the distribution of correct answer choices when writing summative exams, both during their initial answer selection and when transferring their answer letters to the Scantron sheet for correction. Despite the fact that more than half of survey respondents indicated that they do make note of answer patterning during exams and that a series of four to five questions with the same letter for the correct answer would encourage many of them to take a second look at their answer choice, the results pertaining to student outcomes suggest that MCQ randomization, even when it does result in short serial arrays of letter-specific correct answers, does not constitute a distraction capable of adversely influencing student performance.

Dans les très grandes classes de cours de première et deuxième années, l'évaluation sommative se déroule souvent par le biais d'examens comportant des questions à choix multiples afin de pouvoir donner rapidement les résultats aux étudiants. Plusieurs versions de ces examens sont souvent préparées et les questions sont brouillées par ordinateur pour dissuader la tricherie. Lors de la préparation de plusieurs versions d'un examen à choix multiples, l'un des paramètres importants à prendre en considération est que chaque version doit être semblable aux autres pour évaluer équitablement les connaissances des étudiants. Ce projet a pour but d'examiner l'influence possible de l'organisation des réponses correctes sur le choix des réponses des étudiants lors de la préparation de plusieurs versions d'un examen à choix multiples. La question spécifique qui a été posée était de savoir si l'existence d'une série de quatre ou cinq questions à choix multiples consécutives pour lesquelles la même lettre représentait la bonne réponse pouvait avoir une influence préjudiciable sur l'aptitude des étudiants à continuer à choisir la bonne réponse alors qu'ils progressent d'une question à l'autre dans la même série. Les résultats des étudiants qui passent de tels examens ont été comparés aux résultats obtenus quand les étudiants passent des examens dont les questions sont les mêmes mais qui ne comportent pas de telles séries. Ces résultats ont été enrichis par les réponses à une enquête auprès des étudiants pour laquelle les étudiants ont été auto-évalués concernant la question de savoir s'ils avaient remarqué la répartition des réponses correctes parmi les choix multiples quand ils passaient des examens sommatifs, à la fois au départ, quand ils choisissaient leurs réponses, et ensuite quand ils transféraient les lettres correspondant à leurs réponses sur la feuille Scanton pour la correction. Malgré le fait que plus de la moitié des répondants aient indiqué qu'ils ne font pas attention à la structuration des réponses pendant l'examen et qu'une série de quatre ou cinq questions ayant la même lettre pour la bonne réponse pourrait encourager beaucoup d'entre eux à regarder de plus près leur choix de réponse, la conclusion concernant les résultats obtenus par les étudiants suggère que la randomisation des questions à choix multiples, même quand elle aboutit à des séries de

réponses correctes identifiées par la même lettre, ne constitue pas une distraction capable d'influencer négativement le rendement des étudiants.

Keywords
question order, multiple choice exam, scrambling, student outcomes

Cover Page Footnote
I would like to thank Ms. Hannah Gray (University of Ottawa) for her assistance with data collection for this project and the University of Ottawa for providing the opportunity and funding to Ms. Gray to participate in this study via the Undergraduate Research Opportunity Program. I would also like to thank Mr. Amir Hamid (McMaster University) for his assistance with the statistical evaluation of the data. Finally, I am grateful to Dr. Pierre Fortier, University of Ottawa, for creating Clinch, the exam bank used not only for creating multiple exam versions for undergraduate courses in Medicine and Health Sciences at the University of Ottawa but also for tracking question performance each time it is used in an exam.

Summative student evaluation should employ an exam format that supports efficient correction and the timely provision of feedback. However, the large class enrolments associated with first- and second-year undergraduate courses in many university and college disciplines present an important challenge to the satisfaction of these requirements. To both address that need and because this exam format agrees well with the fact-dense content that characterizes many of these introductory courses (including those in the health sciences), multiple choice questions (MCQs) frequently comprise a large proportion of every summative exam (Lowe, 1991; Roediger III & Marsh, 2005; Slade & Dewey, 1983).

The use of MCQ exams is by no means a perfect way of assessing student knowledge and understanding. The development of a reasonably-sized bank of effective MCQs is labour-intensive and MCQs can be criticized for sometimes cueing students to the correct answer because it needs only to be recognized within the list of possible answer choices rather than provided *de novo*. And, on occasion, the correct answer can even be selected via a lucky or strategic guess (Kuechler & Simkin, 2010; Tamir, 1990). But there are also a number of important advantages associated with the use of MCQs for summative assessment. Evaluation is objective and carefully crafted questions can address a variety of cognitive levels. Furthermore, computer grading is accomplished with speed and accuracy and a broad range of curricular content can be assessed within a single summative event (Khan, Tabasum, Mukhtar & Iqbal, 2013; Kuechler & Simkin, 2010; Lowe, 1991). Finally, the ability of an examination, as a whole, to accurately and reliably assess student knowledge and understanding can be determined by the calculation of the level of difficulty and discrimination associated with each MCQ composing that exam, a statistical evaluation that routinely accompanies computer-based exam grading (Lowe, 1991). The difficulty index reveals the percentage of students selecting the correct answer to each MCQ while the discrimination value, by comparing performance on each question between the higher- and lower-scoring students, rates the ability of each MCQ to be answered correctly more frequently by students who achieved higher overall scores on that particular exam (Lowe, 1991; Sevenair & Burkett, 1988).

It must also be recognized that an MCQ exam format is especially vulnerable to cheating, especially when exams are written by large groups of students accommodated under crowded conditions (Khan et al., 2013). In an effort to reduce the ability of students to copy from one another, several exam versions are often created via the process of MCQ scrambling (Bresnock, Graves, & White, 1989; Khan et al., 2013). This scrambling can focus on two possible examination parameters: the arrangement of the individual questions composing that exam can be shuffled and/or the order of the answer choices linked with each question can be rearranged (Bresnock et al., 1989; Khan et al., 2013; Sue, 2009). An important benefit of MCQ reorganization, be it by the order of the questions and/or by the answer choices, is that it makes the summative approach to student evaluation as fair as possible by allowing all students to be assessed using the same questions (Khan et al., 2013; Sue, 2009). As shall be seen, however, MCQ reorganization can be accomplished in many different ways and it is important to be certain that all exam versions encountered by students in a given class evaluate student knowledge fairly and in an equivalent manner and that question order is not unfairly penalizing a particular subset of students (Sue, 2009).

There are a variety of criteria that can be used to organize the individual MCQs comprising an exam. MCQs can be ordered sequentially or by level of difficulty or they can be grouped by topic and a number of studies have investigated a possible influence of MCQ arrangement on student outcomes. Sequential order delivers the questions in the order that the content was covered during lectures whereas reverse sequencing does exactly the opposite. In a chapter contiguity

exam, the questions are grouped by chapter, but within each chapter-related exam section, the questions are not in chronological order, and the chapter-batches themselves do not have to be inserted consecutively into the exam (Balch, 1989). Questions can also be ordered by increasing or decreasing level of difficulty (Noland, Russell & Madden, 2014; Perlini, Lind & Zumbo, 1998). Finally, the advent of computer programs to create exam banks and to permit the random rearrangement of questions has allowed another option, the truly random-order exam, to become increasingly popular (Sue, 2009).

The results obtained from studies comparing sequential versus chapter contiguity versus random exams are controversial. Balch (1989) compared all three approaches to exam construction within a large class of psychology students with the added constraint for the randomly-ordered exam that no two questions from the same chapter could immediately follow one another. His hypothesis was that factual information would be more easily retrieved if accessed in the order that it was first learned and encoded in long-term memory. Indeed, higher-order learning involves the restructuring and organizing of new knowledge so that it can form a part of long-term memory by linking with that which is already known, in this way facilitating subsequent retrieval for application (Kirschner, 2002; Kirschner, Sweller, & Clark, 2006). Balch (1989) reported a small but significant difference in student outcomes between the sequential and chapter contiguity groups and suggested a slight beneficial effect for sequential compared to random. However, the latter effect was noticeable only at a level of $p < 0.10$, a level not usually used as a benchmark for significance. In contrast, a number of other studies comparing sequential and random ordering of questions and involving a variety of disciplines did not find it advantageous for students to answer exam questions in the same order that they had encountered the content during lectures, be they students of large- or small-enrolment classes (Bresnock et al., 1989; Kagundu & Ross, 2015; Khan et al., 2013; Neely, Springston, & McCann, 1994; Sue, 2009; Tal, Akers, & Hodge, 2008).

Other researchers explored question ordering by level of difficulty and also found no significant effect (Laffitte, 1984; Noland et al, 2014; Paretta & Chadwick, 1975; Perlini et al., 1998). An early study conducted by Laffitte (1984) involved 82 undergraduate students studying introductory psychology and compared the ordering of questions in only two ways: by increasing level of difficulty and by random distribution. Perlini et al. (1998) expanded this study, again involving students of introductory psychology, to compare both easy-to-hard and hard-to-easy question ordering with random distribution and also found no difference in student outcomes. An important concession made by the authors was that they could not control the order in which students chose to answer the questions when completing the exam, only the order in which they first encountered them (Perlini et al., 1998). In an even earlier study involving over 300 introductory accounting students, Paretta & Chadwick (1975) attempted to control the order in which students answered MCQs by instructing students to answer the questions in the order presented, suggesting that otherwise they may run out of time to complete the exam. In their study that involved the same three question patterns (easy-to-hard, hard-to-easy, and random), but with the incorporation of the perceived time constraint, they found that average students had poorer outcomes if they encountered the more challenging questions at the beginning of the exam; however, they also reported that weaker students and stronger students had similar examination outcomes regardless of the MCQ pattern. Finally, Noland et al. (2014) revisited the notion of question order by comparing two groups only: those with harder MCQs at the beginning of the exam and those with harder MCQs at the end. They evaluated this approach with several different courses related to accounting, and a variety of class sizes with results that did not always

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

completely agree. However, their overall conclusion was that question ordering by difficulty did not influence examination outcomes (Noland et al., 2014).

This study also explores the influence of individual MCQ ordering with regard to student outcomes, but from a very different perspective. Multiple exam versions were created by computer-based scrambling meaning that for all exam versions the questions were presented in a random order and were not grouped by topic or ordered sequentially. There was no scrambling of the answer choices for each question, only the MCQs themselves were organized so as to be presented in different orders. However, computer-based randomization of question order does occasionally result in exam versions where the same letter-specific correct answer occurs in a series of several consecutive questions. The possible influence that this parameter may have on student approaches to MCQ answering and student outcomes was explored in two ways. The evaluation of student outcomes from such exam versions (looking specifically at student performance on the serial questions) was compared with those from the exam versions written by the rest of the class in which those questions were not in series. In addition, survey data was collected from students in which they self-assessed the extent to which they pay attention to the distribution of their answer choices when writing summative exams, both during their initial selection of their answer and when the answer choices are accumulating and forming a pattern on the Scantron sheet.

Method

This research exploring a possible influence of serial correct answers on the process of exam writing and examination outcomes was approached in two ways. In the first part of the study, data pertaining to the process involved in selecting MCQ answers was collected from a representative class of anatomy and physiology (ANP) students. Also, student outcomes on select ANP exams administered during the previous three years were compared between versions that did contain serial arrays of letter-specific correct answers and versions that did not. The involvement of students and student-related data collection for this research was approved by our university's Human Ethics Committee (File number H09-06-10B). At our university, the anatomy and physiology content is contained within three first-year level ANP courses (ANP1105, ANP1106, and ANP1107). ANP1105 is usually completed during the fall term of first year and, depending on their program of study, many of these students go on to complete ANP1106 and ANP1107 during the winter term of the same academic year. All three ANP courses are supported by supplementary course web sites that provide access to feedback-oriented, textbook-derived practice questions that are primarily, but not exclusively, MCQs. A proportion of each student's final grade (7-8%, depending on the course) is derived from scores earned when completing these open-book formative quizzes at intervals throughout the term. Before each summative exam, students are informed of the examination question distribution, both in terms of types of questions and approximate number of questions devoted to each major content area.

For the first part of the project, 282 students enrolled in the winter term of an undergraduate first-year course in anatomy and physiology (ANP1107) were administered an optional, anonymous survey via Blackboard Learn, their course management platform (Table 1). The survey questions were designed to collect some preliminary demographic information regarding student experience with university-based and anatomy-and-physiology-based exams and then to ascertain their general approach during an exam when recording their answers to MCQs. Final questions in the survey explored to what extent they believed that their selection of answers to MCQs could be

influenced by whether or not the same letter, as their correct answer choice, was starting to appear in a series (Table 1).

Table 1
Online Survey Administered to 282 Students Enrolled in an Undergraduate Course in Anatomy and Physiology (ANP1107)

1) How many years since you graduated from high school?
□ 1 □ 2-4 □ 4-6 □ more than 6

2) How many ANP courses have you completed?
□ 0 □ 1 □ 2 □ more than 2

3) When answering exam questions, do you immediately transfer your answer choice to the Scantron or do you answer all questions on the exam paper first and then transfer to the Scantron?
- □ I transfer them to the Scantron right away.
- □ I answer all the questions in the exam book and then transfer them to Scantron afterwards.
- □ I answer the questions I know on the Scantron right away and then go back a second time to answer the questions that are more difficult.
- □ There is no specific way that I transfer my answers onto the Scantron.

4) Do you pay attention at all to the pattern of answer choices developing on your Scantron sheet?
□ never □ not usually □ sometimes □ often □ always

5) Do you change your mind or take a second look if you see the same answer choice starting to appear in a series?
□ never □ not usually □ sometimes □ often □ always

6) If you do take a second look, after how many of the same letter in a series would you start to pay attention to the pattern and take another look at your answer choices?
□ 3 □ 4 □ 5 □ 6 □ 7

7) If so, does it matter which letter it is?
□ never □ not usually □ sometimes □ often □ always

8) If it does matter which letter it is, please check off the letter(s) that would be most likely to worry you if they appeared in a consecutive series?
□ A □ B □ C □ D □ E □ N/A

9) If you do change your mind/take a second look, do you look at all of them in the series or just the last one or two?
□ I look at all of the questions in the sequence □ Just the last one or two □ Just the first two

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

The second part of the study was a retrospective analysis of the MCQ outcomes from selected anatomy and physiology examinations administered during the previous three years. These examinations consisted of a mix of four-choice and five-choice questions taken from our departmental ANP exam bank with the majority (62%) of questions being five-choice and the remainder (38%) being four-choice. The exam bank has been gradually built over the past several years using a mix of textbook-publisher-supplied MCQs as well as MCQs developed by the author. Given the subject matter of these first-year courses, these MCQs tend to target primarily the lower-order cognitive skills of remembering, understanding and applying, as defined according to Bloom's revised taxonomy (Anderson, Krathwohl, & Bloom, 2001). Indeed, of the 21 MCQs evaluated in the current study, three are classified as level one (remembering), 10 as level two (understanding), and 8 as level three (applying).

When administering exams within these large enrolment courses, the potential risk that students would copy from one another was addressed by randomly re-ordering the MCQs using the scrambling function of our online exam bank so that two to three versions of the same exam could be prepared. When the properties of exams used during the past three years were evaluated, five sets of midterm and final examinations met the criteria for inclusion in this study. One of these exams, Exam 4, was written by the same students who completed the survey described in the first paragraph of the Methods. The total number of MCQs in each of these exams was 44, 64, 43, 59 and 52 for Exams 1, 2, 3, 4 and 5, respectively. For each of the five exam sets, one version of the exam had 4-5 consecutive questions with the same letter choice representing the correct answer (Serial) while the distribution of those same MCQs was very different (Dispersed) within the other exam version(s). Rather than being ordered consecutively, the 4-5 MCQs were scattered throughout the other exam version(s) and were not integrated into a pattern of having the same correct answer appear as part of a sequence.

Prior to each exam session, the exams themselves were shuffled so that distribution to students was entirely random. After exam completion, student outcomes were tabulated and the performance of key questions (difficulty and discrimination) compared between the Serial and Dispersed exam versions. Difficulty was calculated as the number of correct responses divided by the total number of responses. Discrimination was calculated as the difference between the proportions of students in the top and bottom 27% of the class selecting the correct answer (Sevenair & Burkett, 1988). For those examination sets where there were two exam versions with dispersed questions (Exams 3, 4 and 5), the data from both dispersed versions were pooled prior to statistical evaluation. The possible influence of MCQ distribution on the per cent of students selecting the correct answer was assessed using the Pearson Chi-Square Test of Independence (SPSS Version 23). The ability of these questions to distinguish between stronger and weaker students was evaluated by comparing their discrimination values as a function of question distribution using the paired sample t-test (SPSS Version 23). The distributions of final grade outcomes between the subset of ANP1107 students who completed the voluntary survey and the entire ANP1107 class were compared using, once again, the Pearson Chi-Square Test of Independence. For all statistical evaluations, differences were considered significant at $p < 0.05$.

Results

Out of a class of 282 students, 53 elected to complete the survey, a participation rate of 18.8%. With the exception of a final grade of F (below 40%), the respondents represented all levels of achievement in ANP1107 and formed a sub-population of students that was not different (p =

0.277) from the class population as a whole in terms of final course outcomes (see Figure 1). Approximately one-third of respondents were in their first year of postsecondary study and the majority had already completed one (64.2%) or two (24.5%) previous courses in anatomy and physiology.

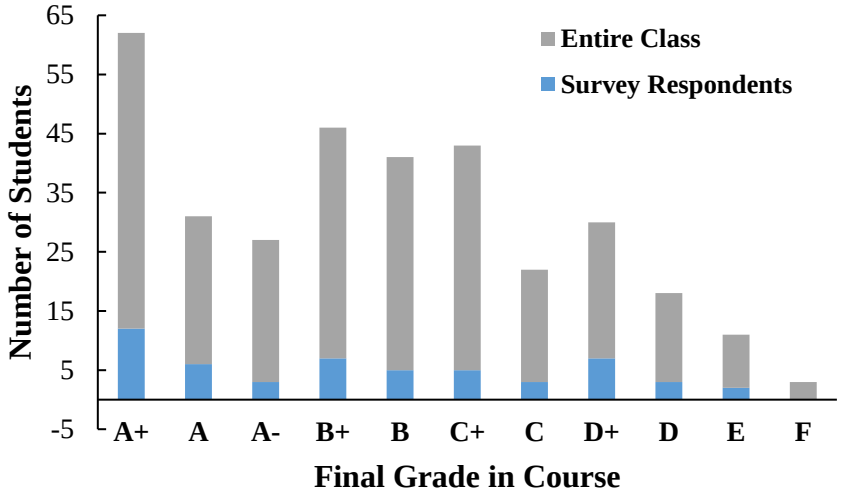

Figure 1. Distribution of the final grades achieved in this undergraduate course in anatomy and physiology (ANP1107) within the entire course (n=282) and among the survey respondents (n=53). It should be noted that, while final grades of E and of F are both failing grades, at the University of Ottawa a final grade of E (40-49%) contributes one point toward a student's grade point average (GPA) and renders a student eligible to write a supplemental exam whereas a grade of F (less than 40%) contributes zero toward the student's GPA and renders that student ineligible to write a supplemental exam.

With regard to survey question 3 (Table 1), all three approaches to completing the Scantron sheet were used by a reasonable number of participants. Immediate transfer, question by question, was the approach used by 32.1% of respondents, whereas 39.6% initially selected and recorded all of their answer choices on the exam pages before batch-transferring them to the Scantron, and 28.3% used a two-step approach in which they answered the questions about which they felt most confident first before returning a second time to tackle those MCQs that were more challenging. Over 60% of respondents indicated that, at least sometimes, they do pay attention to the pattern of answer choices developing on their Scantron sheet and the same proportion indicated a level of inclination to take a second look at their answers if they were starting to form a series (Table 2). For 18.9% of respondents, a series meant three of the same correct answers in a row whereas for the majority of respondents, a pattern became noticeable only if it was four or five in a row (39.6% and 28.3%, respectively).

Table 2
Distribution of Student Responses (Percent of Respondents; n = 53) to Survey Questions 4 and 5

Question	Never	Not usually	Sometimes	Often	Always
4. Do you pay attention at all to the pattern of answer choices developing on your Scantron sheet?	18.9	18.9	45.3	9.4	7.5
5. Do you change your mind or take a second look if you see the same answer choice starting to appear in a series?	13.2	24.5	35.9	22.6	3.8

The identity of the letter (A through E) that was forming the series did not seem important to respondents. In response to question 8 (Table 1), the per cent of students selecting each of the five possible letters between A and E as being particularly worrisome if that letter started to appear in series ranged from a low of 13.2% (B) to a high of 22.6% (C) with the other three letter choices (A, D or E) falling in between those values.

Student outcomes on the MCQs evaluated in this study are summarized in Figures 2 and 3. For Exams 1 and 2 (Figure 2), there were just two versions of each exam (one Serial version and one Dispersed version) and outcomes on each of the four MCQs did not differ ($p > 0.05$; Table 3) between the student populations answering those questions in either of the two exam versions. With regard to Exams 3, 4 and 5 (Figure 3), where there were three versions of each exam, student outcomes were similar for the questions of interest ($p > 0.05$; Table 3) between the Serial MCQ exams and the combined results for the two exam versions in which the MCQs were dispersed. Of particular interest are the two questions shown in Figure 3 to be answered correctly by less than 50% of students (Q2 in Exam 3 [application-based MCQ] and Q3 in Exam 4 [challenging knowledge-based MCQ]). Even for those MCQs, there was no difference in the frequency of correct answer selection by students answering them as part of a series or distributed randomly (Table 3).

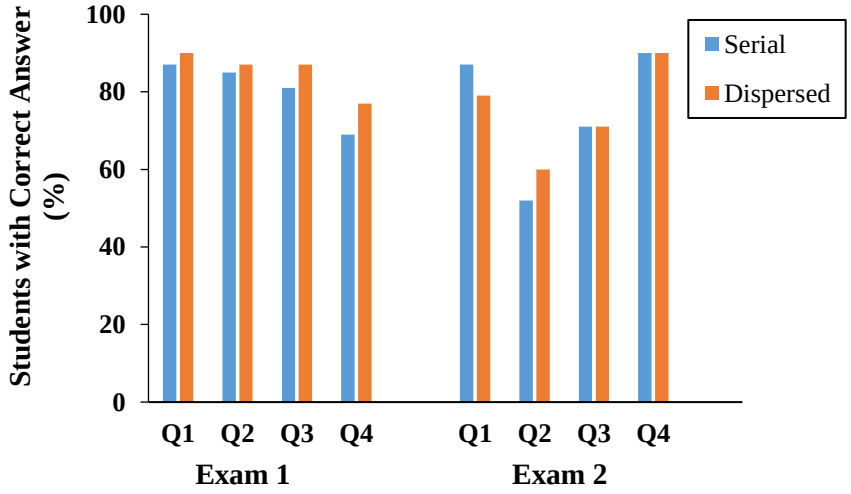

Figure 2. Influence of MCQ correct answer patterning on student selection of correct answer. Results represent two different exams written by students in two different courses. For each exam

there was a version where four MCQs with the same letter-designated correct answer (A for Exam 1, C for Exam 2) were in series (Serial) versus a second version where the same MCQs were randomly scattered throughout the exam (Dispersed). For Exam 1, n = 115 for each exam version; for Exam 2, n = 135 (Serial) and n = 145 (Dispersed).

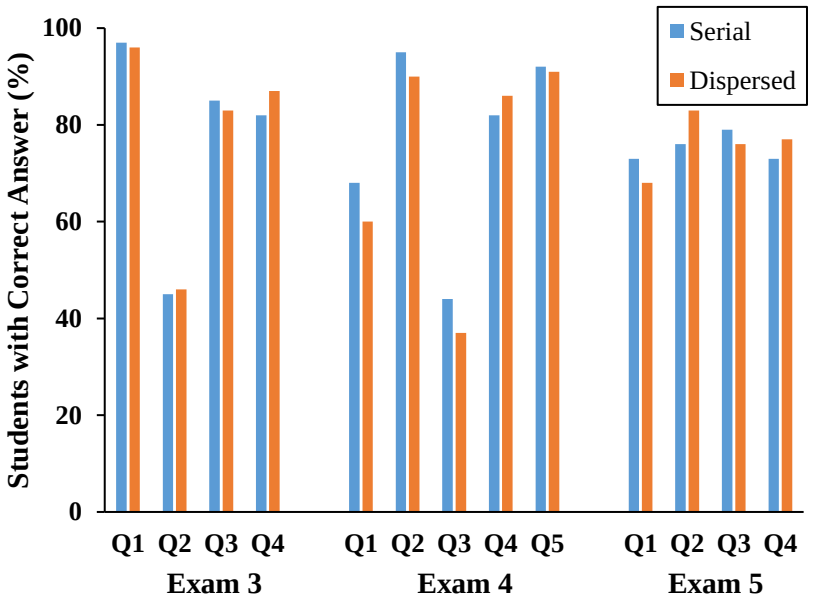

Figure 3. Influence of MCQ correct answer patterning on student selection of correct answer. Results represent three different exams written by students in three different courses. For each exam there was a version where four or five MCQs with the same letter-designated correct answer (C for Exam 3, D for Exam 4 and for Exam 5) were in series (Serial) versus two other versions where the same MCQs were randomly scattered throughout the exam (Dispersed). For Exam 3, n = 78 (Serial) and n = 149 (Dispersed). For Exam 4, n = 91 (Serial) and n = 187 (Dispersed). For Exam 5, n = 97 (Serial) and n = 189 (Dispersed).

Table 3

χ^2 Values Derived from Statistical Comparison of Student Outcomes when Answering MCQs with Same Correct Answer in Series versus Dispersed MCQs

	Exam 1 (N=230)	Exam 2 (N=280)	Exam 3 (N=227)	Exam 4 (N=278)	Exam 5 (N=286)
Question 1	0.378	2.663	0.322	2.005	0.906
Question 2	0.318	1.884	0.043	2.023	1.897
Question 3	1.578	0.000	0.073	1.279	3.372
Question 4	1.371	0.000	1.113	0.463	0.575
Question 5				0.280	

Note. Results from these five different exams involving students in 5 different courses are presented in Figures 2 and 3.

Finally, a comparison of question-specific discrimination values (Table 4) did not reveal a significant effect ($p > 0.05$) of question patterning, whether the series occurred at the beginning,

within the middle region or toward the end of the exam. However, one interesting finding with regard to discrimination pertains to Exam 1 for which the correct answer was choice A (Table 4). Though not significant ($p = 0.077$), there was a consistent trend for the serial questions to be slightly more discriminating, meaning that stronger students seemed routinely more inclined to select A as the correct answer when working with these questions in series, even though those same questions were not answered correctly by a higher total number of students when serially distributed (Figure 2).

Table 4
Discrimination Values for Selected MCQs when Presented in Series (S) or Dispersed throughout the Exam (D)

	Exam 1 (N=230)		Exam 2 (N=280)		Exam 3 (N=227)		Exam 4 (N=278)		Exam 5 (N=286)	
	S	D	S	D	S	D	S	D	S	D
Question 1	0.34	0.22	0.29	0.51	0.11	0.12	0.53	0.48	0.48	0.22
Question 2	0.47	0.31	0.67	0.56	0.56	0.59	0.20	0.25	0.64	0.36
Question 3	0.31	0.25	0.49	0.44	0.16	0.32	0.54	0.24	0.19	0.56
Question 4	0.78	0.56	0.31	0.33	0.53	0.25	0.52	0.33	0.35	0.24
Question 5							0.04	0.27		
F-value	*0.077*		*0.799*		*0.843*		*0.603*		*0.686*	
Correct Letter	A		C		C		D		D	
Series Position	Q31-34 (44)		Q44-47 (64)		Q37-40 (43)		Q51-55 (59)		Q10-13 (52)	

Note. *F*-values are derived from statistical analysis using the paired sample t-test. Student outcome results from these five different exams involving 5 different courses are presented in Figures 2 and 3.

Discussion

This study revealed that student outcomes were not adversely affected by the inclusion of a short series of four to five MCQs with the same letter-specific correct answer within the MCQ portion of an exam. It is not surprising that serialization of the correct answer was not a detriment for those questions that students found to be reasonably easy and that were answered correctly by over 80% of the class. If a student is certain of the correct answer, answer selection will be directed almost exclusively by that student's knowledge of course content and answer patterning should not represent a confounding distraction. The provision of regularly-spaced formative examinations to these students prior to their summative exams may also have allowed them to be well prepared for assessment. Having opportunities to practice retrieving information pertaining to course content when studying is an active study approach that has been suggested to promote learning and long-term retention more effectively than passive processes such as reading the textbook and course notes (Carnegie, 2015; Karpicke, Butler, & Roediger III, 2009; Orr & Foster, 2013; Roediger & Karpicke, 2006). That being said, correct answer position and the possible development of a pattern could become important when a student does not know the correct answer and has resorted to guessing (Attali & Bar-Hillel, 2003). Indeed, Zimmerman and Williams (2003) suggested that guessing can negatively impact the reliability of MCQ examination outcomes, especially if the exams are short and the number of answer choices is small (e.g., true/false questions with a 50% chance of guessing the correct answer). Therefore, it is important to note

that the exams explored in this study were composed of between 43 and 64 MCQs and that each MCQ was associated with a minimum of four and, more often, five answer choices. Furthermore, with regard to the question of whether or not serialization of correct answer choices detrimentally influenced answer selection when questions were more challenging, it should be noted that some of the exams did include within their serial arrays MCQS answered correctly by less than 50% of students. Even those questions were not answered differently by student cohorts writing the Serial versus Dispersed versions of the exam.

Did the identity of the letter forming the series matter? Bresnock et al. (1989) reported that students had better outcomes when writing four-choice MCQ exams if a higher proportion of the correct answers were represented by the letter A rather than the letter D. They postulated that students were more prone to recognize the correct answer if they saw it immediately rather than after reading through several distractors and possibly becoming confused. On the other hand, a pair of studies separated in time by almost 50 years and each controlled to remove knowledge as a basis for answer selection revealed that a higher proportion (70-80%) of study participants, chose central letters (B and C) rather than A or D, the letters at the beginning or the end of the list (Attali & Bar-Hillel, 2003; Berg & Rapaport, 1954). Experimental design (participants not provided with actual answer choice content, only answer letters, and the earlier study also used imaginary questions) assured that participants were purely guessing in order to pick their answer from lists of four options. This tendency, referred to by some researchers as edge aversion and by others as central bias, not only influences guesses made by test-takers, but also guides many of the simple choices we make during our daily lives that are not linked to any sort of a strategic advantage, such as picking a single item from a grocery store display of many identical options (Attali & Bar-Hillel, 2003; Shaw, Bergen, Brown, & Gallagher, 2000).

Interestingly, both central and edge letters were represented as correct answer choices in the exam versions included in this study and, in general, letter position did not exert a significant influence on student outcomes. The majority of the MCQs in the two exams involving serialization of D as the correct answer also included a choice E, so an edge effect with regard to choice D could not be evaluated. However, for choice A, there was one possible instance in which edge aversion may have had a small role to play and that would be for some of the weaker students writing Examination 1. While the total percent of students choosing the correct answer did not differ between the serial and random question arrays, the slight but consistent trend for the serial questions to discriminate between stronger and weaker students, while not statistically significant ($p = 0.077$), suggests that those students who were not certain of the correct answer were somewhat less inclined to select the edge choice, item A, when guessing.

While the survey response rate was low, the fact that this subset of students closely mirrored the class as a whole, in terms of course outcomes, permits increased confidence that the feedback is representative. This low response rate likely derives from such confounding factors as a natural tendency for students to become busier as the term moves toward the time of final exams, the fact that participation was completely voluntary and anonymous, and the need for students to see a direct benefit to their final grade if they are to participate in a course-related activity (Saunders & Gale, 2012).

Interestingly, more than half of survey respondents indicated that they do pay attention to answer patterning on the Scantron sheet and that a series of 4-5 questions with the same letter for the correct answer would encourage many of them to take a second look at their answer choices. However, the results of this study looking at student outcomes suggests that the randomization of MCQ organization is a fair and safe approach to exam creation, even when it does create short

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

series of consecutive MCQs with the same letter-specific correct answer, and that these series did not adversely affect answer choices made by students. It is important to note that the questions were randomized in all exam versions; no students were presented with even the suggestion of an advantage due to sequential examination of course content or gradation of MCQs in order of difficulty. Interestingly, survey data indicated that a proportion of students, possibly close to 30%, may be customizing their MCQ experience by choosing to answer the easier questions first and then going back afterwards to tackle the questions they found to be more difficult, in this way providing them with an easy-to-hard exam approach. As conceded by other investigators (Laffitte, 1984; Perlini et al., 1998), it was not possible in this study to control the order in which students answered the questions, only the order in which they encountered them. While the results of most studies have shown that ordering by difficulty does not confer an advantage on examination outcome (Laffitte, 1984; Noland et al., 2014; Paretta & Chadwick, 1975; Perlini et al., 1998), in this study, that approach may have had the consequence of delaying the recognition of a series by a subset of students because they initially skipped over one or more of the serial questions due to difficulty, only to return later to select an answer and transfer it to the Scantron sheet.

The results of this study suggest that the generation of short series of MCQs with the same letter-specific correct answer does not adversely influence student outcomes. These results should be interpreted with some caution because the MCQs involved in the current study addressed lower order cognitive skills pertaining primarily to knowledge and application and it may be that serial arrays would undermine student confidence in correct answer selection when answering MCQs addressing the higher order cognitive skills of analyzing, evaluating and creating (Anderson et al., 2001). And such MCQs would be found, for example, in examinations for disciplines such as mathematics (students may need to calculate the correct answer) or the social sciences and law (students may need to select the correct justification for a course of action or link appropriate concepts). An additional caveat is that it was also not possible to control the time spent by each student reviewing their answers, noticing letter-specific series, or considering possible changes to their answer selections. It is the author's experience that some students spend very little time reviewing their answers and hand in their completed marking sheets long before the exam time has expired. Other students labour over each answer choice and sometimes end up talking themselves out of a correct answer selection for a variety of reasons that could include the distracting influence of a series. A study involving focus groups and examinations that are not for credit would allow control to be exerted over the time permitted for students to select and transfer their initial answer choices to the computer marking sheet as well as the time allowed to review and possibly revise their answer choices.

In conclusion, the randomization of MCQ order using computer-based exam banks provides an efficient means to create multiple exam versions for a variety of disciplines. While a side effect of this approach may be the creation of short series of MCQs with the same letter-specific correct answer, the results of this investigation suggests that these series do not appear to be an important distraction for students writing these exams.

References

Anderson, L. W., Krathwohl, D. R., Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives*. New York, NY: Longman.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement 40*(2), 109-128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Balch, W. R., (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology, 16*(2), 75-77. http://dx.doi.org/10.1207/s15328023top1602_9

Berg, I. A., & Rapaport, G. M., (1954). Response bias in an unstructured questionnaire. *Journal of Psychology, 38*, 475-481. http://dx.doi.org/10.1080/00223980.1954.9712954

Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: question and response position. *Journal of Economic Education, 20*(3), 239-245. http://dx.doi.org/10.1080/00220485.1989.10844626

Carnegie, J. (2015). Use of feedback-oriented online exercises to help physiology students construct well-organized answers to short-answer questions. *CBE-Life Sciences Education, 14*(3), 1-12. http://dx.doi.org/10.1187/cbe.14-08-0132

Kagundu, P., & Ross, G. (2015). The impact of question order on multiple choice exams on student performance in an unconventional introductory economics course. *Journal for Economic Educators, 15*(1), 19-36.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*(4), 471-479. https://doi.org/10.1080/09658210802647009

Khan, J. S., Tabasum, S., Mukhtar, O., & Iqbal, M. (2013). The effect on student performance of scrambling questions and their stems in medical colleges' admission tests. *Journal of the College of Physicians and Surgeons Pakistan, 23*(12), 904-906.

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10. http://dx.doi.org/10.1016/S0959-4752(01)00014-7

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discover, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86. http://dx.doi.org/10.1207/s15326985ep4102_1

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovated Education, 8*(1), 55-73. http://dx.doi.org/10.1111/j.1540-4609.2009.00243.x

Laffitte, R. G. Jr. (1984). Effects of item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology, 1-1*(4), 212-213. http://dx.doi.org/10.1177/009862838401100405

Lowe, D. (1991). Set a multiple choice question (MCQ) examination. *British Medical Journal 302*, 780-782. http://dx.doi.org/10.1136/bmj.302.6779.780

Neely, D. L., Springston, F. J., & McCann, S. J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology, 21*(1), 44-45. http://dx.doi.org/10.1207/s15328023top2101_10

Noland, T. G., Russell, H. J., Madden, E. K. (2014). Does question order matter? An analysis of multiple choice question order on accounting exams. *Journal of Advances in Business Education, 2*(1), 1-15.

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

Orr, R., & Foster, S. (2013). Increasing student success using online quizzing in introductory (majors) biology. *CBE-Life Sciences Education, 12*(3), 509-514. http://dx.doi.org/10.1187/cbe.12-10-0183

Paretta, R. L., & Chadwick, L. W. (1975). The sequencing of examination questions and its effect on student performance. *The Accounting Review, 50*(3), 595-601.

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology, 39*(4), 299-307. http://dx.doi.org/10.1037/h0086821

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155-1159. http://dx.doi.org/10.1037/0278-7393.31.5.1155

Saunders, F. C, & Gale, A. W. (2012). Digital or didactic: Using learning technology to confront the challenge of large cohort teaching. *British Journal of Educational Technology, 43*(6), 847-858. http://dx.doi.org/10.1111/j.1467-8535.2011.01250.x

Sevenair, J. P., & Burkett, A. R. (1988). Difficulty and discrimination of multiple-choice questions: a counterintuitive result. *Journal of Chemical Education, 65*(5), 441-442. http://dx.doi.org/10.1021/ed065p441

Shaw, J. I., Bergen, J. E., Brown, C. A., & Gallagher, M. E. (2000). Centrality preferences in choices among similar options. *Journal of General Psychology, 127*(2), 157-164. http://dx.doi.org/10.1080/00221300009598575

Slade, P. D., & Dewey, M. E. (1983). Role of grammatical clues in multiple choice questions: An empirical study. *Medical Teacher, 5*(4), 146-148. http://dx.doi.org/10.3109/01421598309146431

Sue, D. L. (2009). The effect of scrambling test questions on student performance in a small class setting. *Journal for Economic Educators, 9*(1), 32-41.

Tal, I. R., Akers, K. G., & Hodge, G. K. (2008). Effect of paper color and question order on exam performance. *Teaching of Psychology, 35*(1), 26-28. http://dx.doi.org/10.1080/00986280701818482

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education, 12*(5), 563-573. http://dx.doi.org/10.1080/0950069900120508

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*(5), 357-371. http://dx.doi.org/10.1177/0146621603254799 .