Abstract

Technology use is increasing in higher education, particularly for test administration. In this study, Capaldi's (1994) sequential theory, which postulates that the specific order of reinforcements and nonreinforcements influences persistence in the face of difficulty or failure, was applied to online multiple choice testing situations in regard to easy and difficult questions. Preliminary data appear to confirm that the order of easy and difficult questions on a test has an impact on student efforts in completing the test. These data may be especially important in the context of online learning, where the teacher is physically absent, as well as in situations where tests are administered through technology-based learning platforms. In all cases, test performance plays a role in student motivation.

AUTHORS

Ronald Mellado Miller, Ph.D.
Utah Valley University

Maureen Snow Andrade, Ph.D.
Utah Valley University

The Effects of Test Question Order on Task Persistence

Increasingly, teaching and learning processes in higher education institutions are technology infused. Technology is a valuable resource for instructors, both in and outside of the classroom, particularly as used within learning management systems. These systems provide learners with access to a range of learning materials and activities and allow faculty to track student participation and progress through various assessments such as assignments and tests (Falvo & Johnson, 2007).

Such systems are also used as the platform for online courses. Over six million students in U. S. higher education institutions are taking at least one online course. CEO survey respondents indicated that their institutions offer approximately 40% of all courses in online or blended delivery modalities, with 67% of these courses taught by full-time faculty (Magda, 2019). These CEOs represent public institutions, which are experiencing a constant increase in online course enrollments (Seaman, Allen, & Seaman, 2018).

Student success, regardless of course delivery modality, is a significant issue for both faculty and administrators. Concerns about lack of progress, typically measured by retention and graduation rates, are paramount. Approximately 61% of incoming students are retained to the second year at their starting institution, while about 73% persist to their second year at any institution (NSC Research Center, 2018). Six-year graduation rates for full-time students obtaining a bachelor's degree are 60% nationally (National Center for Education Statistics, n.d.)

CORRESPONDENCE

Email
maureen.andrade@uvu.edu

Assessment of learning impacts student success. As such, examining how tests are structured and administered and how students manage testing environments is critical. Technology is frequently used across course delivery modes to administer assessments. While instructors may review tests for standards of reliability and validity, a variable that may be overlooked in test construction is that of student persistence in completing assigned tasks. In other words, when faced with difficult test questions or problems, what factors influence a student to make repeated attempts to solve the problem at hand?

Given the increase in numbers of students taking online courses where they are physically separated from an instructor, as well as increased numbers of students taking online assessments for face-to-face courses through a learning management system, this information could be particularly important. The purpose of this study was to examine if the order of easy and difficult questions on a test had an impact on student persistence in completing the test.

Literature Review

While instructors may review tests for standards of reliability and validity, a variable that may be overlooked in test construction is that of student persistence in completing assigned tasks.

Faculty members across higher education institutions are responsible for assessing students' knowledge and skills. A common mechanism for this is formal testing. Faculty typically do not receive training in test construction as part of their PhD work, which is discipline-based, and may therefore rely on publisher-provided tests or create their own with varying degrees of success. Thus, increasing awareness of issues that impact effective test construction is advantageous in terms of obtaining accurate measurements of students' learning and constructing tests in ways that support student persistence in test completion.

A key issue related to testing is the impact of various question order strategies on performance (Bard & Weinstein, 2017; Caudill & Gropper, 1991; Pettijohn & Sacco, 2007; Tal, Akers, & Hodge, 2008). Random or sequential question order appears to have little effect (Cordero, Layson, Martinez, & Quindoza, n.d.; Tal et al., 2008) nor does reverse question order (Pettijohn & Sacco, 2007). Practice prior to an exam accompanied by predictions and postdictions on performance similarly has little impact; however, high achieving students have been found to be more accurate in predicting their performance than low achieving students although the former were underconfident while the latter were overconfident (Bol, Hacker, O'Shea, & Allen., 2005). The order of multiple-choice response items also appears to have little effect on the difficulty of items, although when the correct answer occurs last, the effect on difficulty is slightly increased (Hohensinn & Baghaei, 2017).

Studies have also examined variations such as easy-difficult, randomized, and difficult-easy scenarios (Bard & Weinstein, 2017; Weinstein, & Roediger, 2010, 2012), often in conjunction with student predictions of performance or self-evaluations (Bard & Weinstein, 2017; Bol et al., 2005; Hacker, Bol, & Bahbahani, 2008; Weinstein, & Roediger, 2010, 2012). These studies may involve having students pause between sets of questions to make judgments about their performance. When the easiest questions are presented first, students have more positive perceptions of their performance than when the opposite occurs, These perceptions are maintained throughout the test, perhaps due to students anchoring their positive self-evaluations based on the initial test questions (Weinstein & Roediger, 2010, 2012). This phenomenon also occurs with multiple test attempts—students anchor their self-evaluations based on question difficulty on the initial test attempt and do not adjust them to account for changes in question order on additional test attempts (Bard & Weinstein, 2017).

Capaldi's (1994) sequential theory also provides insights into research on question order and is particularly applicable to online multiple choice testing situations. The theory postulates that the specific order of reinforcements and nonreinforcements influences persistence in the face of difficulty, whether impossible or simply difficult questions. To apply this to testing, a reinforced trial would be one where the question is easy and a nonreinforced question would be a difficult question. For example, when faced with a series of difficult questions followed by an easy question, DDEE (where D indicates a difficult question trial and E an easy one), a student would learn the relationship SD→E, where SD represents the stimulus memory of a difficult problem. Thus, the student would learn that the memory of difficult problems (SD) is followed by easy problems (E).

In the face of many difficult problems, sequential theory seems to suggest that the student will persist in the task in anticipation of the easier problems to come (Capaldi, 1994). If trained in the EEDD sequence, however, the student would learn that SE→D, or that the stimulus memory of easy problems (SE) is followed by difficult problems (D). In this case, when faced with many difficult problems, sequential theory seems to suggest that the student will not persist in the task, as only difficult problems from then on are anticipated (Capaldi, 1994). As such, the student is less motivated to continue. Interestingly, this appears to the most common way for tests and texts to arrange their problem sets, with the easiest problems first and the most difficult problems being presented later.

In addition, sequential theory suggests that the amount of training as the result of test taking (or learning based on the patterns encountered) can have a dramatic effect on the persistence effects discussed earlier. For example, given a series, DEDE, a short test versus a long test should give different results. In the series DEDE, SD→E, where difficult problems lead to easy ones, this pattern would be learned. However, on a short test, the pattern would not be expected to be learned to the extent expected on a long test. Thus, it would be anticipated that a student trained with DEDE would be more persistent when faced with many difficult problems after a longer test than a shorter test.

While other experiments (e.g., Skinner, 1999; Cizek, 1994) have examined the question of order effects on student performance, no study has examined them utilizing the specific predictions given by sequential theory. For example, Skinner (1999) found that students who did well on difficult questions given first, did better on easy questions later on the test than students who had easy questions first. The test takers did not deal with question order on an item by item basis, but were affected by patterns of questions. Cizek (1994) described the results of his experiment dealing with order effects as unpredictable. Others, like Perlini, Lind, and Zumbo, (1998) found that arrangements of test question difficulty had little effect on overall performance. Thus, sequential theory, with its unique item-by-item predictions, appears to offer new insights into the effects of easy and difficult test question arrangements.

Method

The experiment was designed to examine sequential theory in the context of academic technology-based testing. Participants in this study were 38 undergraduate psychology students at a university in the Western United States. Participants were given extra credit for their participation.

Four multiple choice tests were created, an Easy Test (EEEEE), an Alternation Test (EDEDE), an ED Test (EEEDD), and a DDE Test (EDDEE). The tests consisted of either ten (short test) or fifty (long test) questions. An additional Impossible Test of thirty questions was created which consisted of impossible questions for which correct answers were removed and substituted by incorrect answers. Each subject performed one of the regular tests and was then given the Impossible Test. As with most tests, it was anticipated that the students would complete the entire test. Thus, the dependent variable was the amount of time spent trying to solve the Impossible Test questions.

In terms of predictions, it should be noted that each of the tests, the Alternation Test (EDEDE), the ED Test (EEEDD), and the DDE Test (EDDEE), except for the Easy Test (EEEEE), was equated for the percent of easy and difficult questions, with each condition consisting of 60% easy and 40% difficult questions and each condition beginning with an easy question. Only the order of the questions was manipulated. Once a sequence was given (for example, EEEDD), it was then repeated until either 10 trials were given (in the short training) or 50 trials were given (long training). A difference was hypothesized for each of the compared tests regarding time spent solving the Impossible Test questions.

For the Easy Test, it was hypothesized that the long test students would spend less time on the impossible questions as they were only trained S^E→E, which represents that the memory (S^E) of an easy question predicts that another easy question is coming. Therefore, when faced with difficult questions, the student would spend less time trying to solve them, especially in the long test condition where students would be more familiar with the S^E→E pattern. It was thought this would occur because, while trained S^E→E , the memory of difficult

questions, S^D, would not have been conditioned to any stimulus, and so would result in less response time.

For the Alternation Test, it was hypothesized that the training of $S^D{\rightarrow}E$ would lead the long test students to spend more time on the impossible questions than the short test students. This is because the stimulus $S^E{\rightarrow}E$, or the condition that difficult questions would be followed by easy questions, would have been more strongly conditioned through greater training.

For the ED Test, it was hypothesized that the training of $S^E{\rightarrow}E$ would lead to the long test students spending less time on the impossible questions than the short test students as the stimulus $S^E{\rightarrow}D$, that easy questions would be followed by difficult questions, and that difficult questions are only followed by difficult questions.

For the DDE Test, it was hypothesized that, since $S^{2D}{\rightarrow}E$ would become the most persistent in the impossible tasks as they were trained that two difficult questions would be followed by easy questions.

Materials

To form the tests used, a number of questions were extracted from various databases. The databases included the Myers (2003) and the Nairne (2005) introduction to psychology test databanks. The questions were labeled as either easy or difficult as judged by a panel of psychology students. The impossible questions were formed by removing the correct answer from a question and substituting it with an incorrect answer.

Examples of easy and difficult questions follow.

Easy Question:
Deafness refers to the inability to:
> a. walk
>
> b. see
>
> c. hear
>
> d. talk

Difficult Question:
Concept of the id, ego, and superego is best regarded as a theory about 3:
> a. Separate personalities inhibiting one body
>
> b. Different sets of reaction patterns within each personality
>
> c. Separate stages in personality development
>
> d. Distinct types of unconscious conflicts

Procedure

The students were given the web address of the test and were required to finish it in one sitting. The test consisted of the schedule already outlined, given in either two (short test) or ten (long test) sets of five questions. After completing that portion of the experiment, all students were given six sets of five (for a total of thirty) impossible questions. As per Teevan, Zarrillo, & Greenfeld (1983), the dependent variable measured was time, in this case, the amount of time taken to answer each of the impossible questions.

Results

A 2 (Training: Short versus Long) X 4 (Schedule: Easy, Alternation, ED, DDE) X 6 (Sets) X 5 (Questions) repeated measures mixed model analysis of variance (ANOVA) was run with Training and Schedule utilized as between subjects measures and Sets and Questions as the repeated measures. The dependent variable was the time in seconds to answer each of the impossible questions. Descriptive statistics for the study can be found in Table 1. Inferential statistics from the ANOVA can be seen in Table 2. A summary of the results can be seen in Figure 1.

Table 1
Estimated Marginal Means-Training *Schedule

| Schedule | Training | Mean | SE | 95% Confidence Interval | |
				Lower	Upper
Easy	Long	23.6	7.55	8.16	39.0
	Short	27.1	9.80	7.13	47.2
Alternation	Long	38.8	7.17	24.21	53.5
	Short	26.2	11.63	2.46	49.9
ED	Long	25.2	7.55	9.74	40.6
	Short	38.7	8.75	20.81	56.5
DDE	Long	25.3	7.55	9.85	40.7
	Short	41.9	8.75	24.01	59.7

Table 2
Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Sets	24961	5	4992	2.512	0.032
Sets *Training	7664	5	1533	0.771	0.572
Sets *Schedule	62640	15	4176	2.101	0.013
Sets *Training * Schedule	26358	15	1757	0.884	0.583
Residual	298150	150	1988		
Questions	10784	4	2696	1.562	0.189
Questions *Training	3771	4	943	0.546	0.702
Questions *Schedule	38878	12	3240	1.877	0.044
Questions *Training * Schedule	38462	12	3205	1.857	0.047
Residual	207101	120	1726		
Sets *Questions	86747	20	4337	2.738	< .001
Sets *Questions * Training	59130	20	2957	1.866	0.013
Sets *Questions * Schedule	147325	60	2455	1.550	0.007
Sets *Questions * Training *Schedule	148949	60	2482	1.567	0.005
Residual	950601	600	1584		

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Training	6730	1	6730	0.717	0.404
Schedule	10200	3	3400	0.362	0.781
Training *Schedule	28432	3	9477	1.010	0.402
Residual	281447	30	9382		

Note. Type 3 Sums of Squares

Figure 1
Mean Seconds Spent on the Impossible Questions
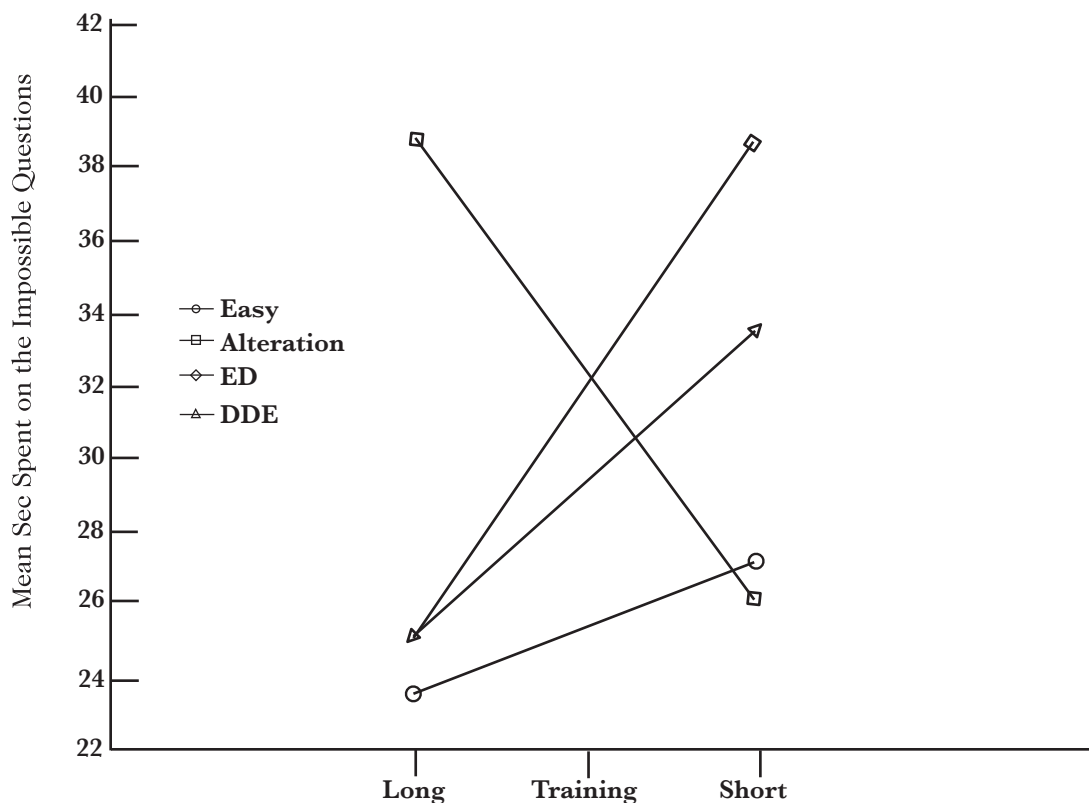
Overall, for the Easy Test, the mean seconds decreased slightly from the short to the long test while the Alternation Test increased from the short to the long test. Additionally, the ED Test and the DDE Test both decreased in their mean seconds spent on the impossible questions from the short to the long tests.

Discussion and Conclusion

In sum, when constructing tests, faculty members should avoid having easy questions followed by difficult questions and should ensure that difficult questions are followed by easy ones.

Overall, all hypotheses appear to have been supported. The Easy Test and the ED Test did decrease in the time spent on the impossible questions, and the Alternation Test did increase time when comparing performance on the short versus the long tests. However, the DDE Test decreased in a way very similar to that of the ED Test.

Thus, the order of easy and difficult questions on a test appears to have an impact on a student's efforts to complete the test, as suggested by sequential theory (Capaldi, 1994). It is important to emphasize that each of the tests was equated for percentage of easy and difficult questions. Additionally, the Alternation Test became the most persistent of all of the tests in the long test condition, presumably resulting from the established pattern of being conditioned to anticipate difficult questions being followed by easy questions. While the DDE test did not perform in the long test as anticipated, this may simply be the result of the long test not being long enough to train the necessary relationship. While further investigation seems to be required, these data seem to support sequential theory and may be especially important in the context of distance learning, where the presence of the teacher is physically absent and the context of given tests may play a greater role in student persistence and motivation.

Additionally, as most textbook-based tests give the easiest questions first and build up to the most difficult, this essentially creates a situation like the Easy Test with easy questions followed by what (to the student) may seem to be impossible questions. It would be better for texts and exams to mix the easy and difficult questions such that difficult questions are followed by easy ones and not vice versa.

While we acknowledge that the sample size for this study is small, another implication is the standard way of giving exams, where either easy and difficult questions are randomly given or where easy ones are generally followed by difficult ones, should be re-examined. Given that the order a student encounters of easy and difficult questions, some students may be given an advantage or be disadvantaged by current practice.

In sum, when constructing tests, faculty members should avoid having easy questions followed by difficult questions and should ensure that difficult questions are followed by easy ones. Pedagogically, faculty members often simply increase the difficulty of problems on a test; however, easy or easier questions should be interspersed with the difficult questions.

In terms of limitations of the study, the number of subjects was small, and a follow-up study with more participants and perhaps other conditions would add clarity. Additionally, this test was not a regular course exam; therefore, the motivations of the students for doing well may have been other than they would be on an exam that counts toward a grade. While this does not seem to have had an effect on the theorized outcomes, it may have played a role in the patterns students focused on and may have enhanced the effects seen here. In spite of these limitations, the study provides practical applications for test construction.

References

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal of Experimental Education, 73*(4), 269–290.

Capaldi, E. J. (1994). The sequential view: From rapidly fading stimulus traces to the organization of memory and the abstract concept of number. *Psychonomic Bulletin & Review, 1*(2), 156-181.

Caudill, S. B., & Gropper, D. M. (1991). Test structure, human capital, and student performance on economics exams. *The Journal of Economic Education, 22*(4), 303–306.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement, 54*, 8-20.

Cordero, H. M., Layson, J. K., Martinez, B. M., & Quindoza, M. P. (n.d.). The effect of question order and paper color on reading comprehension of undergraduate students of UP Diliman: An empirical investigation.

Falvo, D. A., & Johnson, B. F. (2007). The use of learning management systems in the United States. *TechTrends, 51*(2), 40-45.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*(2), 101–121.

Hohensinn, C. & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica, 38*(1), 93-109.

Magda, A. J. (2019). *Online learning at public universities: Recruiting, orienting, and supporting online faculty*. Louisville, KY: The Learning House, Inc.

National Center for Education Statistics. (n.d.). Fast facts. Graduation rates. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=40

NSC Research Center. (2018, June 27). Persistence & retention – 2018. Retrieved from https://nscresearchcenter.org/snapshotreport33-first-year-persistence-and-retention/.

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.

Pettijohn, T. F., & Sacco, M. F. (2007). Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology, 34*(3), 142–149.

Seaman, J. E., Allen, I. E., & Seaman, J. (2018). Grade increase: Tracking distance education in the United States. Oakland, CA: Babson Survey Research Group. http://www.onlinelearningsurvey.com/highered.html

Skinner, N. F. (1999). When the going gets tough the tough get going: Effects of item difficulty on multiple-choice test performance. *North American Journal of Psychology, 1*(1), 79-82.

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.

Teevan, R. C., Zarrillo, D., & Greenfeld, N. (1983). Relationship of fear of failure to amount of time spent in a final examination. *Psychological Reports, 53*(2), 343-346.

Weinstein, Y., & Roediger III, H. L. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition, 38*(3), 366–376.

Weinstein, Y., & Roediger III, H. L. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition, 40*(5), 727–735.