The Use of Traditional Summative Testing to Maximize Student Achievement: An Empirical Study

Ian A. Van Deventer
Northcentral University

Many studies have proven that traditional testing improves the long-term retention of information. Though this growing body of literature is evidence that testing can be used for improving student learning of complex information, very little research sought to determine which form of testing was best for improving student learning. This study attempts to address this gap in the literature using data collected from a sample of students enrolled in first-year financial accounting courses. The specific purpose of this study is to determine whether multiple-choice or short-answer quizzes most significantly improve student achievement in financial accounting.

INTRODUCTION

Studies related to testing-effect theory, the idea that repeated testing improves the subsequent recall of information, are not new. The benefits of testing have been studied for several decades and the studies have evolved significantly since those early days. Though the growing body of literature regarding testing-effect theory supports the use of low-stakes quizzes and high-stakes exams to promote learning of information (Bangert-Drowns et al., 1991; Blunt & Karpicke, 2011; Gates, 1917; Turney, 1931; Spitzer, 1939; & Standlee, 1960), there is little evidence to suggest which form of frequent testing best accomplished achievement of complex information (Gay, 1980). To address this gap in the literature, a study was conducted to investigate the effects of two forms of traditional testing, multiple-choice quizzes and short-answer quizzes, on student achievement in financial accounting. Determining which type of frequent summative testing best supports the learning of accounting information was central to the study.

LITERATURE REVIEW

The most influential theory associated with this study was the testing-effect theory, which was proposed as early as 1917 by Gates, and later tested in 1931 by Turney, Spitzer in 1939, Standlee in 1960, and Bangert-Drowns et al. in 1991. The testing-effect theory was based on the field of human memory and refers to the higher probability that information will be retained and later recalled when at least one quiz or exam was used in the learning process (Bangert-Drowns et al., 1991; Gates, 1917; Spitzer, 1939; Standlee, 1960; Turney, 1931).

In an influential study conducted by Deck (1998), it was determined that students performed better when they were tested more frequently than less frequently. Another important contribution to the testing-effect theory was made by Blunt and Karpicke in 2011. They demonstrated that retrieval practice is a powerful way to promote meaningful learning of complex concepts (Karpicke, 2011; Karpicke &

Grimaldi, 2012). McDaniel et al. (2011) also made an important contribution to the testing-effect theory. The researchers suggested that most other studies were limited because the tests used in those studies were not the summative tests used to evaluate students for the course (Roediger et al., 2010, Roediger, et al., 2011). They determined that students who participated in other studies were not motivated to give their best effort because the tests had no impact on their overall grades.

As a result of these studies, it was recognized that students participating in my study needed to be tested frequently and throughout the financial accounting courses to achieve the maximum benefits from testing. For this reason, students were tested on a weekly basis. The weekly quizzes served as the independent variable and mid-term and final exams were used to measure student achievement, the dependent variable. In order to provide the proper incentive, the quizzes and exams were included in overall grades with the quizzes accounting for approximately 11% of the overall grade and the mid-term and final exams accounting for approximately 8% of the final grade.

METHODOLOGY

The purpose of this study was to determine whether multiple-choice quizzes or short-answer quizzes are better for improving student achievement. The study was conducted in a traditional college classroom setting, specifically first-year financial accounting courses. A quasi-experimental, between-subject research design was used to examine the relationship between quizzes and student achievement. The study was divided into three groups of participants, consisting of two treatment groups and a control group, which corresponded to three different sections of the financial accounting taught by the researcher. One treatment group (section of the course) was administered multiple-choice quizzes. The other treatment group was administered short-answer quizzes. The control group was not administered quizzes. Student achievement was measured using the mean percentage for correct answers on the pre-test, three mid-term exams, and the final exam. To test the difference between the means, the three groups were compared. A mixed-design ANOVA was conducted to facilitate an incremental analysis of student achievement throughout financial accounting.

DATA

The data that are the basis for the study were collected from students who were studying first-year financial accounting. A power analysis was performed and a sample size of 39 participants was estimated for the overall study. The repeated measures, within-between subjects ANOVA was analyzed using G*Power 3. The specific type of power analysis was a priori, which was computed using the following input parameters: an effect size f of 0.25, alpha error probability of 0.05, power of 0.95, 3 groups, 5 measurements, a correlation among repeated measures of 0.5, and a nonsphericity correction of 1. The study included 62 students taking the courses residentially, in a traditional classroom setting. Of the 62 students who were enrolled in the courses, 17 students either chose not to participate in the study or withdrew from the courses and, consequently, the study. The final sample consisted of exams completed by the remaining 45 students.

Table 1 in the appendix reports the mean score for each group by exam. The pre-test was administered primarily as a means of tracking student participation in the study, but the results of the pre-test are interesting and reveal how much students knew about financial accounting at the beginning of the study and how much they learned about financial accounting since taking the pre-test. The mean scores on the pre-test for the control group (n = 15), the multiple-choice group (n = 15), and the short-answer group (n = 15) were 4.73, 5.60, and 5.87, respectively. At this point, the scores are roughly the same. The scores begin to diverge as the mid-term exams are administered. The mean scores on first mid-term exam for the control, the multiple-choice, and the short-answer groups were 59.80, 75.60, and 64.67, respectively. The mean scores indicate that students learned a significant amount about financial accounting since taking the pre-test. It is also evident that students who received multiple-choice quizzes achieved the highest mean score. The mean scores on the second mid-term exam for the control, the multiple-choice, and the short-answer groups were 60.60, 80.20, and 65.60, respectively. The mean scores on the third mid-term exam for the control, the multiple-choice, and the short-answer groups were 63.27, 82.93, and 68.00, respectively. The mean scores on the final exam for the control, the multiple-choice, and the short-answer groups were 62.67, 80.87, and 66.47, respectively.

The within-subjects effects are displayed in Table 2. There was a significant effect of exams, $F(4, 168) = 577.327$, $p < .05$. The interaction results were also significant, $F(8,168) = 4.225$, $p < .05$. Tests of between-subjects effects were also completed. Results are displayed in Table 3. The main effect of group (treatment) was significant, $F(2, 42) = 9.054$, $p < .05$. The between-subjects effects are also plotted in Figure 1 in the results section.

A pairwise comparison was also completed. The results of the comparison are listed in Table 4. There was no significant difference between the mean test scores for no quizzes and the short answer quizzes ($p = .857$) in financial accounting, $p > .05$. There was, however, a significant difference between the mean test scores for no quizzes and multiple-choice quizzes ($p = .001$) and a significant difference between the mean test scores for multiple-choice quizzes and short-answer quizzes ($p = .013$) in financial accounting, $p < .05$. Though there was a noticeable increase in the mean exam scores for students who experienced short-answer quizzes, the mean exam scores were not significantly different from the mean exam scores for no quizzes.

RESULTS

The descriptive statistics in Table 1 indicate that all groups achieved roughly the same mean exam scores for the pre-test. The mean exam scores on the mid-terms and final exam were highest for participants who received the multiple-choice treatment. Also, the mean exam scores for those who received the short-answer treatment and no treatment were close in range with only four or five percentage points separating the groups. Though the mean scores for short-answer and no quizzes were close, the mean scores for short answer quizzes were consistently higher than no quizzes.

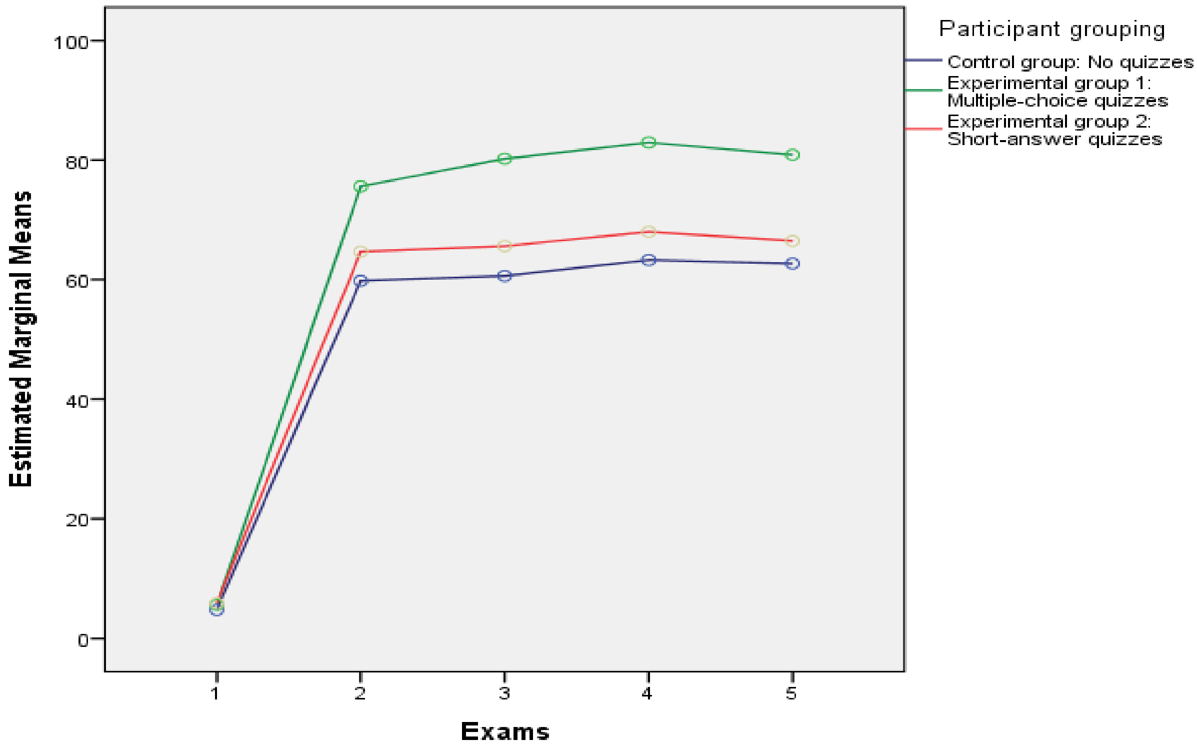
The within-subjects effects are displayed in Table 2. There was a significant effect of exams, $F(4, 168) = 577.327$, $p < .05$. This means that the exam taken significantly affected the score achieved. The interaction results were also significant, $F(8,168) = 4.225$, $p < .05$. Though the mean score was affected by the exam taken, the effect on the score was different in each group.

The between-subjects effects in Table 3 reveal that the main effect of group (treatment) was significant, $F(2, 42) = 9.054$, $p < .05$, meaning that the scores were different across the treatment groups. Recall the discussion of mean scores in Table 1, in which, it was stated that the mean exam scores for mid-term and final exams were highest for participants who received the multiple-choice treatment. Though the mean scores for short-answer and no quizzes were close, the mean scores for short answer quizzes were consistently higher than no quizzes.

The pairwise comparisons in Table 4 indicate that the group receiving no quiz treatment achieved significantly different exam scores than the group receiving the multiple-choice treatment, $p = .001$. The group receiving no quizzes did not achieve statistically different exam scores than the group receiving the short-answer treatment, $p = .857$. Compared to the short-answer treatment, the multiple-choice treatment was again significantly different.

The results of the study indicate that student achievement was significantly improved in financial accounting as a result of multiple-choice quizzes, but not as a result of short-answer quizzes and no quizzes. As a result of this study, accounting educators now have evidence that multiple-choice quizzes can be used to significantly enhance student achievement. See Figure 1 for a graphical representation of student achievement as a result of frequent multiple-choice quizzes. Additionally, the study demonstrates that quizzing can be implemented as low-stakes activities to motivate students. A purpose for using quizzes is to provide students with an opportunity to recall information they previously learned.

FIGURE 1
ESTIMATED MARGINAL MEANS FOR EXAM SCORES



Along the x-axis, the exams are numbered 1 through 5 for the pre-test, the three mid-term exams, and the final exam, respectively. The blue line represents mean exam scores for participants of the control (no-quiz) group. The gray line represents mean exam scores for participants of the short-answer treatment group. The green line represents mean exam scores for participants of the multiple-choice treatment group.

CONCLUSION

This study demonstrates the testing-effect, which is the theory that student learning can be improved by using frequent testing to recall information. In this study, two forms of traditional summative testing were compared. The data collected demonstrate that traditional testing significantly improves the learning of complex accounting information. Specifically, the results show that multiple-choice and short-answer quizzes have a positive effect on student achievement. Though the data show that student achievement was improved when short-answer quizzes were administered to students, it was clear that multiple-choice quizzes had a statistically significant effect on student achievement and short-answer quizzes did not.

REFERENCES

Agarwal, D. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2010). Improving student learning through the use of classroom quizzes: Three years of evidence from the Columbia Middle School project. *Society for Research on Educational Effectiveness.* Abstract retrieved from https://www.sree.org

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research.* Retrieved from http://www.tandfonline.com/toc/vjer20/current

Blunt, J., & Karpicke, J. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science.* Advance online publication. doi:10.1126/science.119937
Deck, D. W. (1998). *The effects of frequency of testing on college students in a principles of marketing course* (Doctoral dissertation, Virginia Polytechnic Institute and State University). Retrieved from http://scholar.lib.vt.edu/theses/available/etd-110298-195932/
Gates, A. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40). Retrieved from http://www.archive.org
Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement, 17*(1), 45-50. Retrieved from http://www.onlinelibrary.wiley.com
Karpicke, J. D. (2011). Tests help students learn. *Curriculum Review, 50*(7), 10-11. Retrieved from http://curriculumreview.com
Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review, 24*, 401-418. doi:10.1007/s10648-012-9202-2
Logan, J. M., Thompson, A. J., & Marshak, D. W. (2011). Testing to enhance retention in human anatomy. *Anatomical Sciences Education, 4*(5), 243-248. doi:10.1002/ase.250
McDaniel, M., Agarwal, P., Huelser, B., McDermott, K., & Roediger, H. (2011). Test-enhanced learning in a middle school classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399-414. doi:10.1037/a0021782
Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*(1), 142-147. doi:10.1037/a0030890
Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2010). *Using testing to improve learning and memory.* Washington University. Retrieved from http://www.apa.org/pubs/journals/xhp/index.aspx
Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom; Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395. doi:10.1037/a0026252
Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641-656. doi:10.1037/h0063404
Standlee, L. S., & Popham, W.J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology, 51*(6), 322-325. Retrieved from http://www.apa.org/pubs/journals/edu/index.aspx
Turney, A. H. (1931). The effect of frequent short objective tests upon the achievement of college students in educational psychology. *School and Society, 33,* 760-762.

APPENDIX

TABLE 1
DESCRIPTIVE STATISTICS

Exams	Treatment Grouping	Mean	Std. Deviation	*N*
Pre-Test	No quizzes	4.73	1.710	15
	Multiple-choice quizzes	5.60	1.844	15
	Short-answer quizzes	5.87	2.416	15
	Total	5.40	2.027	45
Mid-Term Exam I	No quizzes	59.80	17.305	15
	Multiple-choice quizzes	75.60	12.732	15
	Short-answer quizzes	64.67	13.993	15
	Total	66.69	15.931	45
Mid-Term Exam II	No quizzes	60.60	12.982	15
	Multiple-choice quizzes	80.20	13.986	15
	Short-answer quizzes	65.60	12.783	15
	Total	68.80	15.446	45
Mid-Term Exam III	No quizzes	63.27	12.589	15
	Multiple-choice quizzes	82.93	12.887	15
	Short-answer quizzes	68.00	16.767	15
	Total	71.40	16.265	45
Final Exam	No quizzes	62.67	9.484	15
	Multiple-choice quizzes	80.87	12.217	15
	Short-answer quizzes	66.47	13.799	15
	Total	70.00	14.126	45

TABLE 2
TESTS OF WITHIN-SUBJECTS EFFECTS

Source	*df*	*F*	Mean Square	Sig.
Exams	4	577.327	36793.851	.000
Exams * Group	8	4.225	269.264	.000
Error(Exams)	168		(63.731)	

TABLE 3
TESTS OF BETWEEN-SUBJECTS EFFECTS

Source	*df*	*F*	Mean Square	Sig.
Group	2	9.054	4429.231	.001
Error	42		(489.214)	

TABLE 4
PAIRWISE COMPARISONS

Participant Grouping	Participant Grouping	Std. Error	Sig.	Lower Bound	Upper Bound
No quizzes	Multiple-choice	3.612	.001	-23.834	-5.820
	Short-answer	3.612	.857	-12.914	5.100
Multiple-choice	No quizzes	3.612	.001	5.820	23.834
	Short-answer	3.612	.013	1.913	19.927
Short-answer quizzes	No quizzes	3.612	.857	-5.100	12.914
	Multiple-choice	3.612	.013	-19.927	-1.913