Journal of Learning and Teaching in Digital Age, 2022, 7(2), 297-306
https://dergipark.org.tr/en/pub/joltida
ISSN: 2458-8350 (online)

Research Paper

Formative Learning Assessment with Online Quizzing: Comparing Target Performance Grade and Best Performance Grade Approaches

Mark Lubrick*[a], Bill Wellington[b]

[a](ORCID ID: 0000-0003-2420-5150), University of Windsor, Canada, mlubrick@uwindsor.ca
[b](ORCID ID: 0000-0003-3077-4114), University of Windsor, Canada, r87@uwindsor.ca
*Corresponding author

ARTICLE INFO
Received: 14 December 2021
Revised: 03 May 2022
Accepted: 16 May 2022

Keywords:
Formative Learning Assessment
Low-stakes
Online Quizzing
Best Performance
Target Performance

doi: 10.53850/joltida.1036295

ABSTRACT

Background: Using online low stakes MCQ quizzing as a formative learning method is common in many introductory courses; allowing for rapid student feedback. **Purposes:** A comparative study of two different approaches for administering low stakes multiple choice quizzes as tools to enhance formative learning in a large-lecture introductory marketing course was undertaken. **Methodology/Approach:** The sample population was 490 students drawn from two separate cohorts (Fall, n=172; and Winter, n=318). Both cohorts were subjected to 8 sets of quizzes. The Fall cohort's quizzes were scored on the basis of a best performance grade (BPG) while the Winter cohort's were scored employing a target performance grade system (TPG). Learning related outcomes measured included: overall course percentage grades, scores on midterm and final examinations, performance on an alternative exercise, practice exam performance, class participation activity, and time spent on the learning management system (LMS). ANCOVA and MANCOVA analyses were undertaken to compare the two treatments using major, university experience, number of weekly course meetings, number of hours on the LMS, and class participation as covariates. **Findings:** The results indicated that the TPG cohort performed better than the BPG cohort on the final examination and overall course grades. The results were statistically significant. They also had higher first attempt scores on weekly quizzes, though not all results were significant. **Discussion:** The findings indicate that online quizzing scored using a "Targeted Performance Grade" approach is a more beneficial motivation for formative learning than scoring with a "Best Performance Grade" approach.

INTRODUCTION

The idea of using online, low-stakes testing to enhance learning retention via the testing effect (Roediger and Karpicke, 2006), while also acting as a formative learning method (Jacoby et al, 2014) is well established. Roediger and Karpicke (2006, 184-189) report on the "testing effect." They provide strong evidence indicating that repeated testing leads to greater memory retention than repeated study alone. Bell and Cowie (2001) discuss the concepts of summative versus formative learning assessment approaches. Summative learning assessment is designed to monitor student learning and enable a comparison to some learning standards, while formative learning assessment is designed to provide instructors with feedback to monitor teaching effectiveness and allow students to self-monitor their learning (Bell and Cowie, 2001, p. 538).

Brown, Gibbs, and Glover, (2003) present the following list of best practices for formative learning assessments for use in science classes although this list is fully applicable to other disciplines:

- Assessed tasks capture sufficient student time and effort
- These tasks distribute student effort evenly across topics & weeks
- These tasks engage students in productive learning activity
- Assessment communicates clear and high expectations to students
- Sufficient feedback is provided, often enough & in enough detail
- The feedback is provided quickly enough to be useful to students
- Feedback focuses on learning rather than on marks or students
- Feedback is linked to the purpose of the assignment and to criteria
- Feedback is understandable to students, given their sophistication
- Feedback is received by students and attended to
- Feedback is acted upon by students to improve their work or their learning

For use as a formative learning assessment method in introductory courses with large numbers of students, online multiple-choice quizzing (MCQz) presents itself as a highly convenient tool in the modern environment of asynchronous and self-directed learning via electronically accessed Learning Management Systems. The drawback of using online MCQz is that it is hard for this approach to comply with all the requirements of Brown, Gibbs and Glover's (2003) best practices list. For example, the outcome information

of online MCQz is seldom sufficient and detailed enough for effective student feedback. It can also be difficult for instructors to ensure that feedback from online testing is acted upon by students to improve their learning. The student's test performance should ideally be improved by their engagement with productive learning activities. It is also essential that the learning measure is tied to the learning activity. An adaptive release trigger for online quizzes can ensure that learning materials are accessed before testing becomes available and also enable the timing and availability of the quizzes. The authors in this study posited that the use of a minimum target grade of 70 percent (the minimum grade for honors graduation standing in the degree program of the Business School) would encourage students to authentically engage with that material, prior to taking the test and hopefully make it more likely for them to obtain the target grade. Despite the inability of online MCQ Quizzing to satisfy all of the criteria on Brown, Gibbs and Glover's (2003) best practices list, the employment of online MCQz as an approach to formative learning is supported by the conclusions of Jacoby et al (2014). The author's assert that as applied in this study, online MCQz can be capable of meeting or partially meeting the following of Brown, Gibbs, and Glover's, (2003) criteria for formative learning: Assessed tasks capture sufficient student time and effort; These tasks distribute student effort evenly across topics & weeks; Assessment communicates clear and high expectations to students; The feedback is provided quickly enough to be useful to students; Feedback is linked to the purpose of the assignment and to criteria; Feedback is understandable to students, given their sophistication; Feedback is received by students and attended to; Feedback is acted upon by students to improve their work or their learning.

In this study, the authors seek to examine the impact on a course's key summative learning assessment results by using two different learning assessment treatments. The first treatment involved having students acquire knowledge in eight main topic areas through text book readings, online resource materials containing lecture materials, and live lecture class attendance. Their knowledge competency prior to summative assessments could be self-assessed through eight online MCQz topic quizzes. Each quiz was offered with three repetitive attempts and produced three scores. The final quiz grade was based on the "best performance grade" (BPG) result of the three attempts. The second treatment involved the same knowledge acquisition approaches and eight online MCQz topic quizzes. However, in this treatment the quizzes required the achievement of a minimum threshold of performance of 70 percent hereafter referred to as a "target performance grade" (TPG). These quizzes were offered with an adaptive release format and the availability of repetition was tied to achieving the minimum threshold. Once the threshold was achieved, a student could no longer repeat their quizzes. In both treatment instances, one would expect to have the testing effect present. The authors posit that based on the reporting of Bell and Cowie, (2001); Brown, Gibbs, and Glover, (2003); and Jacoby et al (2014); that online quizzing employing an adaptive release format with a target performance grade expectation is more conducive to formative learning than an online quizzing approach based only on repetition. As such, a target performance grade requirement might lead to more learning, and ergo, better performance on summative learning assessments such as the midterm and final exams.

STUDY BACKGROUND

The concept of using testing for formative assessment (Hagstrom 2006) in order to enhance learning retention via the testing effect has a long history in the educational and psychological research literature (Roediger and Karpicke, 2006, 184-189; Tulving 1967). In contrast, the use of online quiz testing as a learning management system tool is relatively new (Angus and Watson, 2009) although its use to assess student learning can now be considered a common practice in education (Gikandi, Morrow and Davis, 2011).

Repetitive testing to enhance knowledge retention involves a number of key considerations, including: the frequency of testing, time spacing between tests, the lag-time between learning/testing, and the learning assessment (Kupper-Tetzel, Erdfelder and Dickhauser, 2014). Wellington, Hutchinson and Faria (2015) studied repetitive testing with online MCQz in a large-lecture style marketing class and reported a beneficial learning outcome as a result. DeSousa and Fleming (2003) report a study that illustrates the "testing effect" and its potential for formative learning in a Theories of Personality course. They compared the examination performance of students who undertook quiz related learning using online MCQz versus in-class MCQz. The students were divided into two different course administration groups where they were examined four times during the course and were required to complete a total of 16 quizzes (four quizzes per exam). The in-class tested students were given paper and pencil quizzes and allowed one attempt for each quiz. The online tested students were allowed to take their quizzes anytime along with up to five repetitions of each of their quizzes with a best grade outcome. DeSousa and Fleming (2003) reported that the examination performance of the online MCQz students was better the in-class MCQz group and that this difference was statistically significant.

Angus and Watson (2009) report on the use of online quizzing in a business mathematics course to enhance learning and performance as measured on examinations. They reported the following: "Our main finding is that exposure to regular (low-mark) online testing significantly improves student learning as measured by a final proctored examination. Importantly, this result is independent of a student's actual performance on each online quiz" (p. 256). In contrast, Andergassen, Modritscher, and Neumann (2014) undertook a study of the effects of repetition on final examination performance on a paper and pencil multiple choice question (MCQ) exam with a sample of 1850 students who were enrolled in three different online courses. They reported weak correlations between repetition of learning exercises and examination performances. The difference in results of these various studies shows the importance of continued research into what aspects of quizzing may impact student examination performance.

Jacoby et al (2014) presented the impact of a "Virtual Learning Environment" (VLE) on Science Education. They set a targeted threshold of performance of 40% on formative learning quizzes as part of eight learning modules in a science course. The students were given access to learning support materials and then had to demonstrate their competency by taking a 10 MCQ online quiz (drawn from a set of 40-60 questions). Students could not progress from one learning module to the next until they had achieved the minimum target level. Jacoby et all (2014) stated that students were given an unlimited amount of time and an unlimited number

of attempts to progress from one module to the next. The students were also presented with two summative assessment tests as they progressed through each module. These summative assessment tests presented 20 MCQs (drawn from the formative assessment question bank) and had 30-minute time limits. The authors reported that course grades increased for students who were part of the new VLE system as compared to students that they had instructed in the course during previous years.

STUDY PURPOSE

The main purpose of this study was to explore the following research question: As part of a formative learning approach applied to a large-lecture introductory marketing class employing eight different MCQz to assess learning in eight topic areas, would students who were asked to attain a targeted level of online MCQz performance (TPG cohort) on each quiz, with up to two additional repeat attempts to achieve the target, outperform students who were simply offered an opportunity of three attempts per quiz performance, with the best performance (BPG cohort) result recorded? The performance outcomes assessed in this study were four-fold: 1) first-attempt scores on each online quiz; 2) final uninflated scores on each online quiz (students in the TPG cohort had their actual scores inflated as a reward if they hit the target sooner than later, only the actual best score was analyzed; only the best score for the BPG cohort was analyzed); 3) midterm examination performance; and 4) final examination performance. A number of secondary performance categories were also examined to see if there were significant differences between the two cohorts. The secondary performance categories include: in-class exercises, online practice midterms, online practice final exams, Assurance of Learning measures, and an optional course exercise; all of which will be explained in detail in the methodology section.

In order to test the research question, the following hypotheses were formulated:

H1: The TPG cohort will have significantly higher course performance scores on their midterm and final examination summative assessments as compared to the BPG cohort.

H2: The TPG cohort will have significantly higher actual first attempt scores on each formative quiz as compared to the first attempt scores of the BPG cohort.

H3: The TPG cohort will have significantly higher final uninflated quiz scores on each formative quiz as compared to the BPG cohort.

H4: The TPG cohort will have significantly higher scores for in-class exercises and for the online practice midterm and online final practice exams, as compared to the BPG cohort.

Hypotheses 1 is based on the notion that the motivation for formative learning associated with a target performance grade quiz assessment will be stronger than that associated with best performance grade assessment, thus leading to more knowledge retention. Hypothesis 2 is formulated based on the idea that the students will be more motivated to learn the material, so they can hit the target grade on their first attempt. Hypothesis 3 is offered on the assumption that target grade performance students will prepare more thoroughly to achieve a higher threshold of performance on their formative quizzes than best performance graded students. H4 is based on the notion that more motivated students are expected to regularly attend classes and thus complete in-class exercises. They should also do better on the online midterm and final examination practice exams. Finally, Hypothesis 4 does not include the secondary performance categories of Assurance of Learning measures, and the optional course exercise because their uptake in the course were highly inconsistent. The reasons for the differential uptake was that the Assurance of Learning measure was only required in the Winter Semester and the Optional exercises were not required at all.

METHODOLOGY

The subjects for this study were first year students drawn from three different sections of a 12-week semester Principles of Marketing mass-lecture style course, delivered to class sizes capped at 250 students, and taught by the same instructor, who used the same textbook and instructional syllabus. The Blackboard LMS was used for communication of information to students, for delivery of key learning support materials, and for the online quizzing which was undertaken.

The course had two major MCQ summative assessments: a midterm exam worth 35% of the course grade and a final exam worth 45% of the course grade. The major examinations were delivered in a traditional in-person, large classroom setting, which was managed by the instructor with proctor/invigilator supports, using scantron sheet exam forms to record and score their answers. The two exams were drawn from question banks and presented in "alternative forms" created by offering the same questions in a scrambled question and scrambled answer form. The final exam also had a quantitative written question involving one of either a breakeven analysis or mark-up pricing problem. As part of regular class meetings, students were asked to take in-class quizzes worth 5% of the course grade. The in-class quizzes were used as formative assessment tools and were composed of pop-up multiple choice and essay questions contained within PowerPoint presentations. These quizzes were graded for completion only and represented a measure of class participation, rather than competency knowledge. Just prior to both the midterm and final exams, the students were required to take an online practice exam to test their readiness. The practice exams represented formative assessments and were each weighted at 1.5% of the course grade and graded on the best performance from three attempts.

The research question was tested as follows: Using the single section cohort of students offered in the Fall semester, the students were required to complete eight online topic quizzes (not to be confused with the online practice exams mentioned above) worth 1.5% of their grade. The quizzes were intended to encourage learning via the testing effect (Roediger and Karpicke, 2006). Each quiz was timed to open after the topic had been presented and to close one week afterwards, offered with three attempts, and graded with a best performance grade approach (BPG). Student feedback was restricted to performance scores only.

In the Winter semester a second cohort of students in two class sections was quizzed using the same pool of online MCQ quiz questions, number of online topic quizzes and LMS delivery platform. Unlike the fall cohort of students, the quizzes were only weighted at 1% each and a target performance grade approach (TPG) using an adaptive release delivery format was employed. This assessment approach would be expected to manifest the testing effect. It was also conceived as representing a superior formative learning approach in accordance with Jacoby et al (2014), because students were asked to meet a minimum threshold of performance of 70% on these quizzes. As soon as they achieved the threshold, the quizzes would shut off.

As an incentive for TPG students to perform well right away, the graded scores of students were inflated in accordance with the number of attempts it took to achieve the 70% threshold. For example: a grade of 7, 8, or 9 on a first attempt would receive a score of 9 and all other scores would be reported as achieved. Consequently, if a student scored 7, 8, 9, or 10, on their first attempt, the quiz would shut off and would report the "raw score" to the students. Then the instructor would review all first attempt grades and then manually inflate scores for those who achieved the threshold. All original quiz scores remained in their raw form on the online learning management system (LMS) quizzing feedback system for students to see. However, official inflated threshold quiz grades were reported separately in the gradebook portion of the LMS. For students who scored less than 7, a second attempt would be available. Second attempts were scored as follows: a grade of 7 or 8 would receive a score of 8 and all other scores would be as achieved. As before, if a student scored 7, 8, 9, or 10, the quiz would shut off and report the "raw score". Once again, the instructor would review all second attempt grades and inflate all scores of 7 to 8 and report current grades in the gradebook. All students who scored less than 7 on this second attempt were offered a third and final attempt to achieve the threshold. On the third attempt, the score would be reported as it was and would not be inflated. Whenever a student took all three quizzes without achieving the target threshold of 7, the best attempt grade was reported.

During both semesters, each quiz was delivered through the Blackboard LMS. Each attempted quiz was made up of 10 multiple choice questions, selected at random from a larger test bank and presented to test knowledge of all the key concepts of marketing presented on the topic. Some of the test banks had to be grouped to ensure full topic coverage but even in groupings, questions were randomly drawn. The test bank sizes were: Quiz 1)– 66 questions; Quiz 2)– 43 questions; Quiz 3)– 51 questions; Quiz 4)– 48 questions; Quiz 5)– 32 questions; Quiz 6)– 40 questions; Quiz 7) – 26 questions; and Quiz 8)– 41 questions. The students were given a pre-set schedule of when the quizzes would be available.

An optional exercise worth an additional 3% of the course grade was available to students in both semesters as part of the University's "Psychology Pool" student research program. This research activity asks students to become subjects in approved research to train Psychology Graduate students and for regular University research involving student subjects. The Psych Pool is voluntary, and students have an opportunity to earn up to 3% added to their course grade by participating. For ethical reasons, all students registered in a course that participates in the Psych Pool have to be offered an alternative, course-based exercise so they can opt out yet have an opportunity to earn these percentage points. The implementation of this optional exercise in the marketing course required the instructor to offer students four grading alternatives as follows: 1) Students need not complete the optional exercise and would be graded on the regular course criteria; 2) Students could participate in the Psych Pool research studies alone and receive credit for a maximum of 3% points added to their grade; 3) Students could do the alternative course-based exercise and receive a grade score that would be worth a maximum of 3% added to their grade; and 4) They could participate in both the Psych Pool and complete the alternative course-based exercise and receive a maximum of 3% points added to their grade, depending on which grade was higher.

For the Fall semester, an online exercise worth 50 marks that tested marketing business acumen was offered as the alternative to the Psych Pool exercise. This exercise was not made available until the 9th week of the semester, closed in the 11th week, and was used as a pilot test for a planned Assurance of Learning (AOL) measure that was intended to be employed in the Winter class for course credit. During the Winter semester, the same set of questions used in the pilot test were made available in the 9th week of the semester as the Psych Pool alternative exercise. In addition, the testing was employed to prepare students for the official AOL test that was administered later. This alternative exercise closed early in the 11th week of the semester.

One confounding occurrence affecting this study was that during the Winter semester, the business school required the instructor to provide an AOL test of the marketing knowledge of the students. This testing procedure was undertaken with the online AOL test piloted in the fall. In the winter class, the testing procedure was worth 4% of the course grade, had a target grade threshold of 70%, allowed for up to four attempts, and was offered via adaptive release. These quizzes operated in the same fashion as the Winter topic quizzes; they closed as soon as a student achieved the minimum threshold. This AOL exercise opened in the 11th week of the semester and then was offered in three stages where it had declining grade weights. If it was completed during the 11th or 12th weeks of the regular semester, it was worth 4%. In order to incentivize early completion a late penalty applied such that if the AOL test was completed after classes ended but before the final examination, it was worth 3%. It was shut down and could not be attempted after the final examination was administered because by rule, the final examination had to be the final knowledge assessment of the course.

The scores of the students from the BPG cohort and the TPG cohort were recorded for all measured class activities. Research Ethics Board approval was obtained to use all aggregate data for this research subject to the requirement that all databases created for analysis were kept confidential by being anonymized and that data security was maintained. The hypotheses were tested by comparing the two treatment cohorts (BPG versus TPG), using their examination performances as dependent variables (Midterm Exam MCQ score, Final Exam MCQ score, Final Exam Written Question Score) by employing both a MANCOVA and separate

ANCOVA analyses, controlling for the variables of business major (1) or non-major (2), number of class meetings (1 for one meeting per week and 2 for two meetings per week), official University semester level classification (Semesters 1 through 9), time spent on the Learning Management System (hours), and Class Participation grades (cumulative grade out of 5). In terms of the semester level classification, 2 students were on exchange from another university and thus, their semester level was unclear, though students in the exchange program are traditionally upper year (3rd or 4th) students. They were categorized as semester 9, while all other students fell within the traditional 1-8 semester levels.

To try and ascertain the nature of student engagement with the course materials over time, the TPG cohort and BPG cohort were compared in terms of the percentage of students in each treatment cohort who completed each exercise. Wellington, Hutchinson and Faria (2015) noted that online quiz completion percentages under a Best Performance Grade (BPG) outcome measure tended to decrease over time, with the first two quizzes having the highest completion rates (94% and 95%) and then rates fell with the later quizzes at 86%, 83%, and 88%, respectively. A similar profile of engagement could be expected in this study for the BPG cohort but if a target grade (TPG) performance system is more engaging a different profile might emerge for this cohort. As such, the course activity completion rate of the two cohorts is measured and reported on. In addition to exercise completion percentages, the amount of time spent on the Learning Management System could be expected to represent a sign of student engagement. However, given the fact that it is uncertain how much time would be spent on the learning management system, no hypothesis is offered. Students in the TPG cohort might be expected to have "fewer" quiz attempts since they would be cut off as soon as they reached the minimum target. Conversely, they might peruse the learning support materials of the LMS longer in order to learn more to achieve the targets. In contrast, the BPG cohort of students might be expected to use more quiz attempts and thus, spend more time on the LMS. However, BPG students might not prepare as well for the quizzes, knowing they would just need the "best" result and thus, use the "testing" alone as their learning approach. The time on the LMS was used as a covariate that might affect performance. Finally, the frequency of class attendance could be taken as a symbol of the level of engagement with the course material. The use of in-class completion quizzes is a proxy measure for class attendance. As such, it is expected that student's who attend class more often would be more engaged with the material and would perform much better on examinations than student's who attended class less. If the TPG performance system is more engaging, this cohort might be expected to attend classes more often and it would be hypothesized that they would have a higher in-class quiz percentage grade. As such, in-class quiz completion was used as a covariate that might affect performance on the midterm and final examinations.

FINDINGS

The MANCOVA and ANCOVA analyses of performance in the course assessment measures between cohorts are reported on in Tables 1,2, 3 and 4. For H1, a MANCOVA comparison of midterm and final examination performances was undertaken with the Midterm MCQ scores, Final Exam MCQ scores and Final Exam Written Question scores as multiple dependent variables and online quiz cohort (TPG or BPG) as the treatment variable. The analysis was undertaken using hours on the LMS, class participation, semester level, university major, and number of weekly class meetings as covariates. Before undertaking the MANCOVA analysis the fitness of the data for this analysis was investigated. It was found that the assumption of equality of covariance between the treatment groups was satisfactory (Boxes M value 7.455, F-value 1.233, Sig. .286) and the assumption of homogeneity of variance for the dependent variables across the groups was also satisfactory (Levene's Test of Quality of Error Variances; Midterm MCQ, F-Value 0.534, Sig. .465; Final Exam MCQ, F-value 1.103, Sig. .294; and Final Exam Written Question, F-value 2.508, Sig. .114). As such, the analysis was completed and interpreted. The overall MANCOVA multivariate test results produced a Pillai's trace value of .053, an F-Value of 9.049, and a Sig. < .001. These results indicate that there was a significant difference in the overall examination performance results between the two cohorts when controlling for the impact of the covariates of hours on the LMS, class participation, semester level, university major, and number of weekly class meetings. In examining the outcomes of the separate ANCOVA results it was found that although the TPG cohort had higher midterm scores than the BPG group (M= 43.17 SD =9.10 vs M= 41.52 SD = 8.98), the difference was not statistically significant (F = .73, Sig. at .394). In terms of the final examination, the TPG cohort had statistically significant higher final exam MCQ scores than the BPG cohort (Mean of 64.98 vs 57.65, F = 18.74, Sig. at <.001). Finally, a comparison of final examination Written Question scores between the two groups indicated the TPG cohort had higher scores than the BPG cohort (Mean of 5.06 vs 4.33) but the results were only marginally significant (F = 3.51, Sig. at .062). The overall findings reported on in Table 1 support the acceptance of H1.

Table 1. MANCOVA Results of Examination Performance of a Target Performance Graded Student Cohort versus a Best Performance Graded Student Cohort Using Semester Level, Major, Number of Class Meetings, Hours on the LMS and Class Participation as Covariates

	Target Performance Graded Cohort			**Best Performance Graded Cohort**			**MANCOVA between Subjects Effects**	
Performance Measure	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**F**	**Sig.**
Midterm MCQ (Max 65)	318	43.17	9.10	172	41.52	8.98	.73	.394
Final Exam MCQ (Max 93)	318	64.97	14.35	172	57.65	13.73	18.74	.001**
Final Exam Written Question (Max 7)	318	5.06	1.79	172	4.33	2.02	3.51	.062*
Multivariate Test Results								
	Value	F	Hypothesis df		Error df		Sig.	
Pillai's trace	.053	9.049	3.000		481		<.001	

** Significant P < .05
* Marginally Significant P < .10

The influence of covariates on the relationship between the different treatment cohorts and the dependent variables of examination performances is reported on in Table 2. The findings in Table 2 indicate that for the covariates of Time spent on the LMS and the Student's semester level, there was an impact on both Midterm Exam MCQ and Final Exam MCQ grades but not on Final Exam Written Question Grades. The covariates of Class Participation and Program Major had impacts on all three examination elements. Finally, the impact of the covariate of number of class meetings per week was shown to not have any direct influence on the relationship between treatment group and examination performances.

Table 2. MANCOVA Results of Impact of Covariates of Semester Level, Major, Number of Class Meetings, Hours on the LMS and Class Participation on Examination Results

	Tests of Between Subjects Effects				
Covariates vs Dependent Variables	**Type III Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Value**	**Sig.**
Hours Spent on LMS					
Midterm MCQ Grade	746.339	1	746.339	9.865	.002**
Final MCQ Grade	2368.837	1	2368.837	13.107	<.001**
Final Written Question Grade	3.628	1	3.628	1.139	.286
Class Participation (Max 5)					
Midterm MCQ Grade	632.411	1	632.411	8.359	.004
Final MCQ Grade	1147.202	1	1147.202	6.347	.012**
Final Written Question Grade	67.236	1	67.236	21.117	<.001**
Semester Level					
Midterm MCQ Grade	1428.893	1	1428.893	18.886	<.001**
Final MCQ Grade	2485.527	1	2485.527	13.752	<.001**
Final Written Question Grade	3.267	1	3.267	1.026	.312
Program Major					
Midterm MCQ Grade	485.633	1	485.633	6.419	.012**
Final MCQ Grade	2404.760	1	2404.760	13.305	<.001**
Final Written Question Grade	47.803	1	47.803	15.014	<.001**
Number of Class Meetings					
Midterm MCQ Grade	.802	1	.802	.011	.918
Final MCQ Grade	329.600	1	329.600	1.824	.178
Final Written Question Grade	2.294	1	2.294	.720	.396

** Significant P < .05
* Marginally Significant P < .10

Table 3. ANCOVA Comparison of Means of Performance Measures Using Semester Level, Major, Number of Class Meetings, Hours on the LMS and Class Participation as Covariates

	Target Performance Cohort			**Best Performance Cohort**			**ANOVA Results**	
Performance Measure	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**F**	**Sig.**
Midterm MCQ (Max 65)	318	43.17	9.10	172	41.52	8.98	0.729	.394
Final Exam MCQ (Max 93)	318	64.97	14.35	172	57.65	13.73	18.741	.000**
Final Written Quest (Max 7)	318	5.06	1.79	172	4.33	2.02	3.508	.062
Hours Spent on LMS (a)	318	16.14	10.78	172	18.00	12.35	8.389	.004**
Class Participation (Max 5) (b)	318	3.49	1.44	172	2.99	1.44	17.185	.000**
Course % Grade (Max 100)	318	70.95	12.80	172	65.76	14.09	5.797	.016**
Alternative Exercise (Max 3)	318	1.76	1.32	172	1.39	1.27	4.452	.035**
Practice Midterm (Max 25)	252	18.52	3.26	129	19.35	3.26	4.501	.035**
Practice Final (Max 25)	235	17.37	3.72	121	16.50	4.22	2.028	.155
Chapter Quizzes (Max 80)	318	53.24	12.84	172	50.05	17.23	0.088	.767
Mean number of quiz attempts	318	1.65	0.43	172	2.39	0.50	223.092	000**
+AOL Exercise (Max 50)	318	38.23	5.91	114	29.35	8.90	117.876	.000**

(a) Analyzed without Hours on LMS as a covariate.
(b) Analyzed without Class Participation as a covariate.

** Significant P < .05
+ Exercise was mandatory for Target Performance Cohort and Optional for Best Performance Cohort

With respect to H2, a comparison of grades on first attempts on quizzes, the findings reported on in Table 4 indicate that for every one of the eight online quizzes, the TPG cohort outperformed the BPG cohort on each first attempt; however, the results were significant for 3 quizzes and only marginally significant for 2 others. These findings partially support the acceptance of H2.
With respect to H3, a comparison of the maximum uninflated quiz scores from all three attempts on each of the eight topic quizzes is reported on in Table 4. The BPG cohort had higher mean quiz performance scores than the TPG cohort on all quizzes, except for Quizzes 6 and 8; however, the results were significant for quizzes 1, 5, and 6, and only marginally significant for quiz 7. As such, the findings support the opposite of what was hypothesized and thus, this hypothesis is rejected.

Table 4. ANCOVA Comparison of Online Quiz Means By Attempt Number Using Semester Level, Major, Number of Class Meetings, Hours on the LMS and Class Participation as Covariates

Quiz # and Grade	**Target Performance Cohort**			**Best Performance Cohort**			**ANCOVA Results**	
	N	**Mean**	**SD**	**N**	**Mean**	**SD**	**F**	**Sig.**
Quiz 1 Max Grade	309	7.43	1.36	158	7.80	1.44	4.775	.029**
Quiz 2 Max Grade	307	7.32	1.41	152	7.55	1.46	1.419	.234
Quiz 3 Max Grade	299	7.24	1.22	141	7.33	1.49	1.585	.209
Quiz 4 Max Grade	298	7.73	1.33	140	7.88	1.41	1.354	.245
Quiz 5 Max Grade	249	7.70	1.24	145	7.90	1.59	6.898	.009**
Quiz 6 Max Grade	278	7.47	1.45	135	7.06	1.76	4.685	.031**
Quiz 7 Max Grade	288	7.46	1.46	130	7.68	1.64	3.708	.055*
Quiz 8 Max Grade	242	7.81	1.39	129	7.71	1.58	1.361	.244
Quiz 1 Attempt 1	309	6.25	1.90	158	5.72	2.08	5.270	.022**
Quiz 2 Attempt 1	307	6.28	1.90	152	5.76	2.07	4.054	.045**
Quiz 3 Attempt 1	299	6.21	1.71	141	5.65	1.93	3.139	.077*
Quiz 4 Attempt 1	298	6.83	1.92	140	6.54	1.91	1.934	.165
Quiz 5 Attempt 1	249	6.73	1.83	145	6.37	1.84	0.575	.449
Quiz 6 Attempt 1	278	6.22	2.04	135	5.67	1.85	6.062	.014**
Quiz 7 Attempt 1	282	6.41	1.93	131	5.87	2.11	3.277	.071*
Quiz 8 Attempt 1	242	6.91	1.92	129	6.55	2.01	1.231	.268

** Significant P < .05
* Marginally Significant P < .10

Table 5. Pearson's R Correlations of Chapter Quiz Performance and Exam Performance by Assessment Cohort

Performance Measures	**Target Performance Cohort N= 318**			**Best Performance Cohort N = 172**		
	V1.	V2.	V3.	V1.	V2.	V3.
V1. Chapter Quiz MCQ Grade	-	-	-	-	-	-
V2. Midterm Exam MCQ Grade	.275**		-	.375**	-	.-
V3. Final Exam MCQ Grade	.359**	.672**	-	.339**	.717**	-
V4. Final Written Question Grade	.362**	.220**	.280**	.364**	.301**	.280**

** R value significance P < .001

In regards to H4, the results in Table 3 indicate that the TPG cohort outperformed the BPG cohort on the in-class quizzes, the online practice midterm, and the online practice final exam, though the results were only significant for the in-class quizzes and online practice midterm. These results generally support the acceptance of H4.

The relationships between chapter quiz MCQ performance grades and examination performance grades were assessed with a correlation analysis between total chapter quiz MCQ scores and exam performance scores. The findings of this analysis are reported on in Table 5. These results indicate that for the BPG cohort, total chapter quiz scores have a significant and medium relationship with examination performance scores in all cases. For the TPG cohort the relationship between total chapter quiz scores and exam performances was weak for the midterm performance (< .30) but medium for the final exam performance. A correlation between Midterm MCQ scores and Final Exam MCQ scores indicated strong correlations between the Midterm and Final Examination performances for both cohorts. However, the Final Written Question grades were weakly correlated with Final Exam MCQ grades for both cohorts. This was the case for the Midterm MCQ grades and the Final Written Question grade for the Target Performance cohort while the Best Performance cohort fell into the lower range of the medium strength relationship classification. Cohen and Cohen, (1983, p. 61), reports that correlations greater than .50 are considered strong, those between .3 and .5 are considered medium, and those between .1 and .30 are considered weak. Finally, the exercise completion rates (in the order in which they occurred) of both cohorts for the online exercises and in-class exercises were calculated as an estimate of engagement. These results are reported on in Figure 1 and the profile reported is very similar to that reported by Wellington, Hutchinson and Faria (2015). Over time, the completion percentages of exercises tended to fall for both the TPG and BPG cohorts. However, with the exception of Quiz 5, the

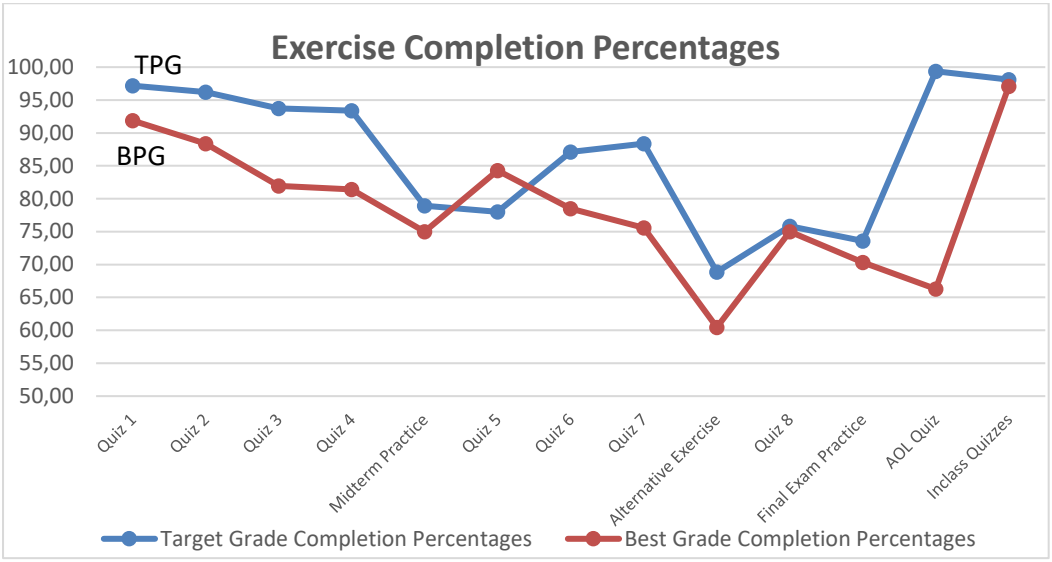

Figure 1. Graph showing exercise completion

TPG cohort had completion rates that were consistently above those of the BPG cohort for all exercises, indicating that there seemed to be a greater level of engagement in the exercises for this group. A further reinforcement to indicate the TPG cohort was more engaged are the ANCOVA findings reported on in Table 3 where the average score for class participation of the TPG cohort is higher than that of the BPG cohort and the difference is statistically significant. TPG students were more diligent in attending classes than the BPG students.

In order to understand if quiz participation rates affected quiz performances an evaluation of grade performances by exercise was undertaken (see Figure 2). Although participation rates fell over time, the average level of performance on the chapter quizzes for those who completed them remained relatively stable and the BPG cohort generally outperformed the TPG cohort for the maximum grade on the raw scores. As discussed earlier in the results section, this was counter to what was hypothesized, yet upon reflection, it is not surprising given the BPG cohort had a stronger incentive to attempt additional quizzes because they were not differentially rewarded with an inflated grade for early success. Further, they were not prevented from repeating quizzes to seek the maximum quiz score of "10". The TPG cohort could only have a second chance at a score of "10" if they initially scored 6 or less. They could top out at less than 10 and were even prevented from having any further attempts to earn a perfect score in this case. As reported on in Table 3, the BPG cohort did have more average attempts on quizzes than the TPG group and the difference was statistically significant (M=2.39 SD=0.50 vs M=1.65 SD=0.43, Sig. < .001).

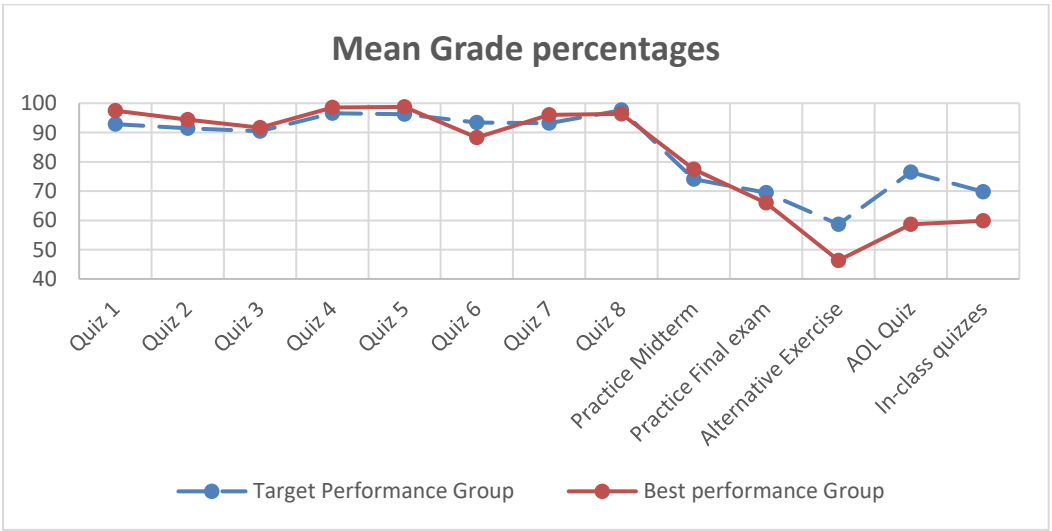

Figure 2. Graph showing mean grades

DISCUSSION AND CONCLUSIONS

The research reported here sought to explore whether students who were exposed to an online testing system with a target performance grade outcome motivation (TPG) would outperform students who were exposed to a best performance grade outcome motivation (BPG) on midterm and final exams. The MANCOVA findings indicate that the TPG cohort of students outperformed the BPG cohort in the three main summative measures of course performance (Midterm MCQ exam, Final MCQ exam, Final Written Question), though the difference in midterm grades is not significant.

In originally designing the study in light of past experiences with online quizzing performances using a Best Performance Grade system, the authors considered that it was likely that students would have attempted quizzes when they were not prepared at all and then studied in response to improve their scores. In contrast, the authors hypothesized that students motivated by a Target Performance Grade performance system would be incentivized to be better prepared when they attempted their quizzes for the first time because they could obtain an inflated performance grade for crossing the minimum threshold of 70%. As such, the quality of their first performance mattered far more. The findings of the study support this, as the TPG cohort outperformed the BPG cohort on first quiz attempts for all of the quizzes, though the differences were not always significant. This situation was to be expected, given that the TPG cohort had their number of attempts limited by the experimental design.

One basic intent of the online quizzing for both cohorts was to encourage the retention of learning. The notion was that there would be a higher quality of formative learning associated with a Target Performance motivation system as opposed to having more testing effect exposures, a feature associated with the Best Performance motivation approach. The correlation results in Table 5 between quiz performances and examination performances was of medium strength for both cohorts bringing into question whether knowledge retention was occurring. This finding is not dissimilar to those reported by Andergassen, Modritscher, and Neumann (2014) and Wellington, Hutchinson and Faria (2015). In the current study, the multiple-choice quizzes used the same types of questions as the midterm and final examinations. As such, the weakness of the relationship could not be due to the difficulty of the questioning. Wellington, Hutchinson and Faria (2015) offered a number of explanations for the lack of correlation between examination and online quiz performances which might apply to this study. They discussed that students would likely access relevant information sources and likely even collaborate together while attempting quizzes. They would be prohibited from doing these things in examination situations. As such, average quiz performances would produce inflated grades for most students and likely not reflect their individual ability to retain knowledge. This study does have findings that provide some evidence that quizzing is affecting knowledge retention though. The TPG cohort performed far better on the final examination MCQ's than the BPG cohort and the difference was highly significant. The level of knowledge needing to be retained for better performance on the final examination would be much greater than the level of knowledge required for the midterm examination which occurred much earlier in the course. It is possible that the impact of better learning retention amongst the TPG cohort simply took longer to manifest itself and this explains why it wasn't until the final examination that a significant difference between the performances of the two cohorts was found.

The authors were concerned that a fatigue factor or boredom factor would set in given the quiz repetition would result in a large number of tests. The assignment completion percentages for all chapter quizzes for the TPG cohort ranged from a high 97.16% on the first quiz falling down to 73.58% on the final examination practice quiz (see Figure 1). The trend line shows a consistent fall in completion percentages over time as the number of online quizzes increased. By comparison, the chapter quiz completion percentages ranged from a high of 91.86% on the first quiz down to 70.35% on the final examination practice quiz for the BPG cohort. Here again, completion rates showed a steady decline from beginning to end. Curiously, the completion percentages for the midterm and final exam practice quizzes for both the TPG and BPG cohorts were very similar (Midterm of 79.25% versus 75%; Final Exam 73.9% versus 70.35%). The in-class exercise completion rates were quite similar, with 98.11% of the TPG cohort completing at least one exercise compared to 97.09% of the BPG cohort completing at least one exercise. The participation rate in the optional exercise was also similar between the two cohorts, with 68.87% of the TPG cohort completing the exercise compared to 60.47% of the BPG cohort. The profile of the BPG cohort for quiz completion percentages was highly reminiscent of that reported by Wellington, Hutchinson and Faria (2015). Conversely, the TPG cohort had consistently higher completion rates for the chapter quizzes as compared to the midterm and final examination practice quizzes, which were based on best grade performance criteria. The TPG cohort participated more in all class activities than the BPG cohort did. Interestingly, both cohorts demonstrated serious drop-offs in participation rates for the midterm and final online practice tests, even though they were required exercises. This is even more curious given that for the TPG cohort, the grade weightings of these assignments were actually higher than their weekly chapter quizzes. It can only be speculated upon, but given students had taken and received feedback from the chapter quizzes throughout the semester, perhaps they did not feel the need to repeat the process in preparation for their midterm and final examinations and/or valued other study techniques more. It is also likely that they were under pressure to prepare for examinations in other subject areas at the same time.

This study has a number of very important limitations. Firstly, the study did not include a completely untreated control group that took the same course section in the same semester from the same instructor with the same examinations. As such, a direct comparison of those students who were learning with and without repeated testing cannot be made. Secondly, although a few students enrolled in the second semester were repeating the class from the fall, the instructor was unable to directly account for them in the study because only aggregate data could be used. It was considered that their potential bias in the findings would be insignificant because there were less than five such students. Thirdly, the study design did not randomly assign students to the two treatment cohorts. This would have produced a more rigorous measure of the impact of different self-assessment motivations on examination performance. However, this kind of treatment is constrained by ethical considerations because within course section and between course sections, the evaluation methods have to be the same during any particular semester of instruction. Therefore, to develop a control group, the course would have to be changed between semesters. This kind of situation could be addressed by involving multiple institutions and multiple classes at the same time. However, this would be offset by introducing more confounding variables, starting with different instructors, different syllabi and textbooks, different lengths of semesters, etc. A more reasonable approach might be to conduct replication studies at other institutions to improve external validity and generalizability rather than seeking more "control".

Despite these limitations, the findings of this study were not dissimilar to those of past studies (Angus and Watson, 2009; Jacoby et al, 2014; and Wellington, Hutchinson and Faria, 2015). Based on the findings in the current study, the authors assert that the use of online testing as a formative approach to learning and retention of knowledge in large-lecture introductory marketing courses is encouraged. When employed, the authors recommend the use of a target performance grade assessment approach as being more effective than a best grade performance assessment approach to encourage formative learning and enhance learning retention.

Ethics and Consent: Ethics committee approval was received for this study from ethics board (REB) at University of Windsor, Social and Human Sciences Research Ethics Committee. The formal permission was under REB# 17-191: "Comparison of two sections of a business course utilizing different weekly quizzing methods".

REFERENCES

Andergassen, M., Mödritscher, F., & Neumann, G. (2014). Practice and repetition during exam preparation in blended learning courses: Correlations with learning results. *Journal of Learning Analytics*, *1*(1), 48–74. https://doi.org/10.18608/jla.2014.11.4

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the Numerical Sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255–272. https://doi.org/10.1111/j.1467-8535.2008.00916.x

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536–553. https://doi.org/10.1002/sce.1022

Brown, E., Gibbs, G., & Glover, C. (2003). Evaluation tools for investigating the impact of assessment regimes on student learning. *Bioscience Education,* 2(1), 1-7. https://doi.org/10.3108/beej.2003.02000006

Cohen, J., & P Cohen. (1983). *Applied multiple regression/correlation analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

DeSouza, E., & Fleming, M. (2003). A comparison of in-class and online quizzes on student exam performance. *Journal of Computing in Higher Education*, *14*(2), 121–134. https://doi.org/10.1007/bf02940941

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Hagstrom, F. (2006). Formative learning and assessment. *Communication Disorders Quarterly*, *28*(1), 24–36. https://doi.org/10.1177/15257401060280010301

Jacoby, C., Heugh, S., Bax, C., & Branford-White, C. (2014). Enhancing learning through formative assessment. *Innovations in Education and Teaching International*, *51*(1), 72–83. https://doi.org/10.1080/14703297.2013.771970

Küpper-Tetzel, Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*(3), 373–388. https://doi.org/10.1007/s11251-013-9285-2

Roediger, & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Tulving. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175–184. https://doi.org/10.1016/S0022-5371(67)80092-6

Wellington, W., Hutchinson, D., & Faria, A. J. (2015). The effectiveness of online quizzing as a repetitive learning tool in a marketing class: A field study of the testing effect. *Developments in Business Simulation and Experiential Learning*, *42*. 42-49. https://absel-ojs-ttu.tdl.org/absel/index.php/absel/article/view/2907