Multiple Attempts for Online Assessments in an Operations Management Course: An Exploration

Ryan K. Orchard

Multiple attempts for online assessments in an Operations Management course: an exploration

ABSTRACT

In Learning Management Systems (LMS), tools for online homework assessments include a number of alternatives for the assessment settings, including the ability to permit students to attempt an assessment multiple times, with options for how the multiple-attempts are administered. A specific implementation of online assessments in an introductory Operations Management course, where students were allowed up to two attempts, found that assessment performance was better for students who attempted only once, even when comparing between students of similar ability in the course. A discussion of why this may be, and how minor modifications to the specific assessment settings could change student behaviors, is provided.

KEYWORDS

online assessments; multiple attempts; operations management; online homework

INTRODUCTION

The use of Learning Management Systems (LMS) to supplement face-to-face courses is generally considered commonplace in Colleges and Universities (Piña, 2013). Most LMS include assessment tools that allow for a myriad of question types and assessment settings and options, such as the ability to allow students to attempt an assessment multiple times, with students being provided some amount of feedback after an individual attempt and with different questions being chosen from a question pool in any subsequent attempt(s). Little research has been done to investigate how different specific assessment characteristics (such as length and question types) and settings (such as the number of attempts permitted, grading, and the amount and timing of feedback provided between attempts) affect student behavior and performance. The following provides descriptive statistics and an exploration of the results of a specific implementation of online assessments in an introductory Operations Management course in the Fall of 2015, including an analysis of which student groups (in terms of grade categories) are more apt to make use of multiple attempts, and how the specific assessment settings may have influenced student behaviors.

BACKGROUND

Homework assignments

Although it seems a given that homework assignments are an important contributor to student success, particularly for quantitative courses, this is not without evidence to the contrary, although perhaps none that is resounding. Fish (2013) studied the *perceptions* that students have of online homework in Introduction to Operations Management courses and points out in their literature review that "the debate [on the effectiveness of homework on student performance]

continues to be mixed." (p. 59). Peters, Kethley, and Bullington (2002) examined the use of homework in an introductory Operations Management course and found that the student group for which homework grades counted toward the overall course grade actually had poorer performance on exams than did the student group for which homework was not graded; however, the former group were aware that the exams were in part directly based on the homework, and their exams would have accounted for more weight, in terms of the course grade, and thus their better performance on these exams may not be completely surprising. The authors also point out that the difference in exam performance was "primarily attributed to performance on non-quantitative exam questions" (p. 343), presumably since homework questions only included quantitative questions.

For our Operations Management course we assumed that graded homework questions *do* contribute to student learning, and we included quantitative questions as well as conceptual/theoretical questions (for example, about the relationship between variables or how a change to a key assumption would affect a model).

Online assessments

Palocsay and Stevens (2008) compared different modes for delivering homework in a quantitative course (introductory Economics) – the traditional textbook-problem approach, and three different web-based approaches - and found no statistically-significant differences in effectiveness between these types of homework systems. Maclean and McKeown (2013), through experimentation, concluded that while formative assessments are indeed effective, the medium (on-paper vs. online) does not seem to matter. Olczak (2013) found a positive relationship between student use of use of a web-based resource for delivering multiple-choice quizzes and exam performance.

For our course, we were comfortable that delivering homework assessments online was as effective in promoting student learning as would be paper-based homework assessments, and we also valued that online assessments provide quicker feedback for students and reduce grading for Instructors, allowing for more value-added time to prepare classroom material. (For additional benefits of online homework over paper-and-pencil methods, see Fish, 2013, p. 59.) We also supplemented the somewhat mechanical and impersonal nature of the online assessments by requiring a few hand-in assignments that were more conceptual and were graded manually. Jungic, Kent, and Menz (2012) support the practice of using *both* online and paper-and-pencil assessments, and point out how the latter can "allow instructors to ask students to prove …demonstrate …[and] graph" (p. 9).

Allowing multiple attempts

Research on the use of multiple attempts for online assessments and the relative effectiveness of different assessment settings is generally sparse and inconclusive. Rhodes and Sarbaum (2015) found that allowing multiple attempts for online homework in their introductory Macroeconomic

classes improved scores, but with evidence of reduced effort, due to the prevalence of guessing for the sake of making use of information provided on a first attempt to improve grades in subsequent attempts. (It is of note that this study used identical questions between attempts and took the best of each students' attempt scores; the authors recommended changes to these conditions that could improve student effort – these will be discussed later.) Stratling (2015) provides general support for the use of "repeat testing" as a way to encourage students to reconcile their understanding (or lack thereof) of concepts with their lecture notes and with their performance on previous tests, and the results of their surveys are that their students strongly agreed that this was helpful in developing understanding of course concepts. Yourstone, Kraye, and Albaum (2010) conducted a study in an undergraduate introductory Operations Management course where they allowed one group of students up to two attempts for online quantitative assessments and another group up to four attempts - the students who were allowed only two attempts performed better, in terms of the amount of improvement between entrance exams and course exams. The authors reason that those students saw the need to be more careful during their limited attempts, which contributed to better learning.

Inspired by these findings, we allowed multiple attempts but capped our attempt limit at two. An additional reason for allowing multiple attempts in our course was that many students seem caught off-guard by our conceptual questions and the overall nature of our assessments, so allowing only one attempt at an assessment that was meant to promote learning seemed too constraining and also often resulted in complaints about low grades and unexpectedly complex questions.

ONLINE ASSESSMENTS IN MGTS 352, FALL 2015

The course

MGTS 352 is an introductory Operations Management course, taught face-to-face to primarily third-year students in class sizes of forty, as a required part of the Bachelor of Commerce program at MacEwan University in Edmonton, Alberta. Following an introductory chapter that discusses Business Strategy in an Operations Management context, the Fall 2015 course covered five main topics, each of which had an associated online assessment: Forecasting, Project Management, Constraint Management, Inventory Management, and Waiting Lines Management. The course was quantitative in nature and included weekly fifty-minute Excel lab sessions in addition to regular classroom (lecture) time. The course emphasized problem-solving and interpretation as much as rote calculation using formulas; the evaluation methods, including online assessments administered using an LMS, were intended to reflect this philosophy.

The assessments

Students were evaluated in a number of ways, including two midterms and a final exam, a case study analysis (completed in groups), multiple (3-5) hand-in homework assignments that were manually graded by the Instructor, and five online assessments. The online assessments

accounted for 20% of the final course grade for each student. Online assessments were administered by way of a "Test" in Blackboard Learn (Release 9.1, April 2014). (From this point forward, the term "test" will be used to refer to an online assessment, to keep with the terminology of the LMS that was used).

Each individual test question seen by a given student was randomly drawn from a question set; for example, the Forecasting test was comprised of 14 question sets where each individual question set consisted of two-to-six questions that were based on the same concept. So, a student's test of fourteen questions would draw these questions from fourteen the corresponding question sets, thus it was very unlikely that any two students would have the exact same questions on their tests, or that two attempts for any individual student would have the exact same questions, although there could be similarities.

Question format was primarily multiple-choice. All test questions were created so that they could be graded by the LMS without requiring any manual grading by the course Instructor, and were edited and vetted over a number of years. Questions varied in difficulty, with some simply requiring straight-forward calculations, some requiring higher-level thinking and problem-solving, and some requiring that the student develop a spreadsheet model to solve them. In general, a test would be made up of an equal proportion of simple, medium, and difficult questions. By design, these tests were by no means simple, nor could they be completed in a short amount of time, and students were warned to be proactive and not to attempt the tests in one sitting on the due date.

Implementation – test options

A test allows for a number of different test options (Blackboard Help, 2016). The following summarizes the relevant options used in our course:

- *Force Completion?* No. (We did not require students to complete the test in one sitting, but rather allowed them to save partially-completed tests and return to them as many times as they wished, up to the due date/time or until they clicked 'Submit'.)
- *Attempts Allowed:* Up to 2.
- *Score attempts using*…last graded attempt. (If a student attempted a test twice, the second attempt was used for their grade. The software allows for other options, such as using the best attempt or an average of all attempts, but in the spirit of allowing multiple attempts to promote concept mastery, we used the last graded attempt.)
- *Show results and feedback to students…*
    - After each/any submission, students were immediately shown their score per question (i.e., correct or not), but not what the correct answers were. (Recall that a student would likely see different versions of the questions should they attempt the test again).

o   After the due date/time, students were shown their score per question, the correct answer for each question, and "feedback" (explanations that were saved into the LMS by the question author at the time that the questions were created.)

DATA

For each student test attempt, Blackboard Learn recorded the grade, date/time submitted, and the duration (time) from when the student first opened they test until they submitted. This attempt information was downloaded for all attempts over the five tests of 146 consenting students; in total, the data used in the current analysis includes 822 test attempts. To account for the general differences in students' academic abilities in the course, which in turn may influence a student's propensity to attempt a test multiple times, we grouped/classified students in terms of final exam grades, in ranges of 10% (e.g., students with final exam scores in the 80%-89.99% range were grouped together), so that during our analysis we would be comparing students of similar academic ability.

OBSERVATIONS AND ANALYSIS

Multiple attempts, by assessment
Information about the individual tests and the proportion of test-takers that attempted each more than once are provided in Table 1.

Test Chapter	Questions	Excel Questions	Total Attempts	% who attempted twice	Avg Grade
Forecasting	14	4	163	30%	74.2
Project Mgt	12	-	167	34%	80.4
Constraint Mgt	3	2	171	41%	81.5
Inventory Mgt	10	-	173	44%	75.2
Waiting Lines Mgt	10	-	148	33%	73

Table 1: summary of test and result information

A regression model was run to examine the relationship between independent variables (the number of questions on a test, whether or not the test included any Excel-based questions, and the relative difficulty of the test based on average grade) and the dependent variable (the proportion of students who took the test that attempted it twice). The result was a relatively insignificant R^2-value of 53%; thus, it seems that none of these factors explain, with statistical significance, the difference in the proportion of students who attempted a given test twice (i.e., we found no statistically-significant linkages between the differences in our five tests and the frequency of multiple-attempts.)

Users of multiple attempts
Overall, of 602 student-tests, 220 (approximately 36.5%) had a second attempt. It is important to note that the nature of the tests in this course (quantitative, some with Excel questions, and

generally requiring some work) as well as test settings (e.g., questions are randomly chosen from question sets and grades are based on second attempts) will be factors in a student's propensity to attempt a test a second time. Descriptive statistics shown in Table 2 suggest that students at the lower end of the grade scale (based on final exam grades) appeared to be more likely to employ multiple attempts when completing online tests; however, a simple chi-square test failed to reject the null hypothesis that the proportion of students in each grade category that attempt tests multiple times are equal (chi-square test value was 0.85). To restate – we did not find statistically-significant differences in frequency of multiple-attempt-taking between student groups.

Final Exam Grade Category	Count of students	Count of tests	Multiple-attempt tests	Proportion of tests attempted twice
<= 49.99%	16	55	20	36%
50% - 59.99%	13	49	21	43%
60% - 69.99%	24	97	36	37%
70% - 79.99%	34	135	50	37%
80% - 89.99%	42	187	69	37%
90% - 99.99%	17	79	24	30%

Table 2: frequency of multiple test attempts, by final exam grade category

Results of multiple attempts
The most interesting observations came from comparing test scores for single attempts with test scores for multiple attempts, for each test and student grouping, as per Table 3. For example, of the students who had a final exam grade between 80% and 89.99%, the average test grade for *single attempts* was 85.2%, while the average test grade for the *first of two* attempts was 65.8% and for the *second of two* attempts the average was 75.9%. In general, first attempts for those attempting a given test twice tended to have very low scores relative to first attempts of those who only attempted once (about 20% less, on average), which may seem obvious since students with a lower first attempt would be more inclined to attempt the test a second time; however, the average scores of *second* attempts were also generally lower than the results for those that attempted only once, even when comparing students in the same final exam grade category. While the results of the second of two attempts may not be much lower than single attempts in most cases, it is of note that students attempting tests twice only outperformed students attempting tests once in one of the grade categories (90%+), and by a very small margin.

		Single Attempt			Multiple - 1st of 2			Multiple - 2nd of 2		
		Count	Avg	StDev	Count	Avg	StDev	Count	Avg	StDev
	Overall	382	79.6%	18.5%	220	59.0%	21.0%	220	72.2%	24.9%
by Test	Forecasting	87	73.8%	16.8%	38	63.8%	16.1%	38	75.0%	18.7%
	Project Mgt	83	81.9%	15.5%	42	62.8%	17.3%	42	77.5%	17.2%
	Constraint Mgt	71	90.1%	19.8%	50	44.0%	24.7%	50	69.3%	36.2%
	Inventory Mgt	67	77.7%	17.4%	53	68.7%	14.8%	53	72.0%	16.6%
	Waiting Lines Mgt	74	75.7%	19.0%	37	56.2%	21.0%	37	67.6%	28.5%
by Grade Category (based on final exam)	20% - 29.99%	4	76.0%	29.9%	1	30.0%	0.0%	1	10.0%	0.0%
	30% - 39.99%	4	65.4%	5.0%	1	67.5%	0.0%	1	70.0%	0.0%
	40% - 49.99%	27	64.4%	22.8%	18	45.9%	26.1%	18	53.9%	29.8%
	50% - 59.99%	28	71.4%	23.0%	21	49.4%	19.5%	21	59.4%	24.3%
	60% - 69.99%	61	75.6%	19.5%	36	53.7%	22.0%	36	68.0%	30.2%
	70% - 79.99%	85	75.7%	17.5%	50	56.6%	19.1%	50	73.9%	19.3%
	80% - 89.99%	118	85.2%	14.2%	69	65.8%	18.3%	69	75.9%	21.8%
	90% - 99.99%	55	91.3%	9.9%	24	71.7%	16.2%	24	92.2%	8.4%

Table 3: single-attempt vs. multiple-attempt results, by test and by final exam grade category

In order to test the statistical significance of these results, two-sample t-tests (unequal variance) on the difference in means were performed, which is similar to the methodology used by Rhodes and Sarbaum (2015). The null hypothesis was that there is *no difference in mean test scores* between single attempts (i.e., using result 1 of 1) and multiple attempts (using result 2 of 2), and we test three different alternative hypothesis: the first (H_1: Diff > 0) asserting that average performance on attempt 1 of 1 is better than average performance on attempt 2 of 2, the second (H_1: Diff $\neq$ 0) that the average scores using attempt 1 of 1 vs. attempt 2 of 2 are not equal, and the third alternative hypothesis (H_1: Diff < 0) that students who attempted two times performed better, on average, than those attempting once. The results are provided in Table 4 and indicate that when comparing between students in similar final exam grade categories, we reject the null hypothesis (no difference between means) in only a few cases, when using a 5% significance value. In both of these cases (the 50% - 59.99% grade category and 80% - 89.99% grade category), it is found with statistical significance that single attempt students outperformed multiple-attempt students. Overall, the result that seems most profound is that *two attempts does not outperform one attempt in any category* – in other words, the alternative hypothesis that average scores for attempt 2 of 2 is greater than average scores for attempt 1 of 1 was resoundingly rejected in all grade categories.

	Diff	SE	d.f.	H$_1$: Diff>0	H$_1$: Diff.≠0	H$_1$: Diff<0
Overall	7.4%	1.9%	359	0.000	0.000	1.000
Forecasting	-1.1%	3.5%	64	0.623	0.753	0.377
Project Mgt	4.4%	3.2%	74	0.084	0.168	0.916
Constraint Mgt	20.8%	5.6%	69	0.000	0.000	1.000
Inventory Mgt	5.7%	3.1%	113	0.036	0.072	0.964
Waiting Lines Mgt	8.1%	5.2%	52	0.062	0.124	0.938
40% - 49.99%	10.5%	8.3%	29	0.107	0.213	0.893
50% - 59.99%	12.0%	6.9%	41	0.044	0.087	0.956
60% - 69.99%	7.6%	5.6%	52	0.091	0.183	0.909
70% - 79.99%	1.8%	3.3%	94	0.292	0.583	0.708
80% - 89.99%	9.3%	2.9%	102	0.001	0.002	0.999
90% - 99.99%	-0.9%	2.2%	51	0.656	0.688	0.344

Table 4: t-test analysis on difference in mean scores for single vs. multiple test attempts

DISCUSSION

The purpose of the current analysis was to examine how specific assessment settings influence student behavior that fosters learning, while bearing in mind the specific nature and conditions of our course and online assessments. In the end, the results of our analysis were indeed both unexpected and noteworthy to us, and have resulted in modifications to the way that multiple attempts are administered in our course.

A first observation is that the frequency of multiple-attempt use (36.5%) was less than expected, likely due to (1) the specific nature of the tests (somewhat difficult and time consuming and sometimes including spreadsheet work), (2) the fact that a second attempt would include different question versions, and (3) that the grade was based on the second attempt, if submitted. The overall weight that these tests had on the course grade (20% of final course grade) may have been a factor, as well. It was also observed that the propensity to attempt assessments multiple times did not, at a statistically-significant level, depend on the general academic ability (in this course) of the student.

The most noteworthy observation from our results is that, in general, average scores on the second of two attempts are not better than average scores on single attempts for any given test or student group; in fact, average two-attempt results are worse than average single-attempt results in most cases. This is not consistent with Rhodes & Sarbaum (2015) – they observed that two attempts result in better scores than single attempts – but quite possibly because of the difference in test settings - they use the same questions on the second attempt as the first, and take the higher of the two scores. Meanwhile, they find strong evidence that "[multiple-attempt] students employ an optimizing behavior [with the goal of achieving] higher homework grades with no additional effort" (p. 130), which is similar to our observation that although many students may plan to complete the tests in a single attempt, with effort and care, and others attempt a test a second time perhaps only because of a lower-than-desired result on a first attempt, there also seems to be a number of students that use a first attempt strategically, as a "throw-away" attempt

to get information that will help on a subsequent attempt. This is also similar to the "observed experimentation by students" that Yourstone, Kraye, and Albaum (2010) describe (p. 349). Perhaps, by allowing two attempts with only the second attempt counting, we enabled or even encouraged the "throw-away first attempt" strategy as a way to get a decent grade without ever having to engage in any arduous thinking. Mortimer Adler (1948), in his essay *Invitation to the Pain of Learning*, wrote that

> "…the fundamental activity that is involved in every kind of genuine learning is intellectual activity, the activity generally known as thinking...without thinking, the kind of learning which transforms a mind, gives it new insights, enlightens it, deepens understanding, elevates the spirit simply cannot occur.
>
> Anyone who has done any thinking, even a little bit, knows that it is painful. It is hard work…it is fatiguing, not refreshing…if allowed to follow the path of least resistance, no one would ever think." (pp. 360-361).

This seems evident in our results as presented previously in Table 3, as well as in the fact that 47 of 220 (21.4%) second attempts actually experienced a grade *decrease* from the corresponding first attempt (i.e. some students didn't engage in arduous thinking on *either* attempt). Whether or not the students who are using multiple attempts to avoid having to engage in deep and effortful thinking on our tests would otherwise be willing to bear down on a single attempt is not clear, but it does seem as though allowing a first "throw-away" attempt might not contribute to learning in the way that was originally hoped. This might not mean that two attempts aren't still useful, though; as Rhodes & Sarbaum (2015) point out (p. 130), perhaps *averaging* the scores of the two attempts (rather than taking the second or the best score) would deter the experimentation approach. In the end, we would like to increase the willingness of a student to roll up their sleeves and put in some effortful thinking on these tests, whether they do them once or twice, and it seems possible that test settings related to multiple attempts might be a way to influence (positively or negatively) this.

Our observations should also be held up to the Stratling (2015) "repeat testing" study, which looked at the effectiveness of testing students on a topic in three successive ways - first by an audience response system ("clicker") tests in a lecture, then by an out-of-class online test that could be repeated if a grade of less than 70% was achieved, and finally by a clicker test in a subsequent lecture. Despite some key implementation differences compared to our testing, and also considering that the Stratling tests did not directly count toward course grades (ours counted for 20% of the course grade), there are still a few insights from Stratling that can be related to our analysis. Stratling found that students who favour a deep approach (DA) to learning felt that the tests had a "greater influence…on their learning motivation and behavior than students who preferred a surface approach (SA) to learning" (p. 1.), which might seem to contradict our conclusions. However, this does not at all imply that DA students would be more apt to strategize to take the same online test multiple times. Stratling describes that the DA students

considered the three different test types to be more beneficial than the SA students did (p. 13), but not that the DA students were taking individual online tests multiple times, although if the DA students did perform at less than 70% they reported being more likely to do "more revision of lecture material before retaking the test" (p. 11). It could thus be argued that the DA students that Stratling found to perceive tests as beneficial learning tools are the same types of students who bore down and submitted one strong attempt for our online tests, while the SA students of Stratling that perceived repeat testing to be less impactful on their learning motivation and behaviour are the same types of students that attempted our tests multiple times either because their surface approach on the first attempt yielded unsatisfactory results, or because the availability of multiple attempts was seen as an opportunity to not have to extend beyond their preferred surface approach on any attempt, which ended up leading to poor results in both attempts since the tests were not trivial.

In closing, as a result of our study, our online tests have since been modified such that (1) the student grade will be based on an average of the two attempts (if they attempt twice), and (2) the students will be provided with a short presentation about the results of our study and Adler's quotation.

Other opportunities for future research on this topic seem to abound. Of immediate interest for our course is to see how student behavior changes with our different test settings. Further analysis into how shorter assessments or a different course (e.g. a non-quantitative course) affects student multiple-attempt behavior would also be of interest. Perhaps surveying the students to better understand when/why they chose to attempt an assessment a second time and to differentiate between tendencies of DA and SA learners could provide useful insights. And, of course, if one could correlate specific assessment settings with impact on overall *learning*, that would be of the most value.

REFERENCES

Blackboard Help (May 10, 2016). Retrieved from https://en-us.help.blackboard.com/

Adler, M.J. (1941). Invitation to the Pain of Learning. *Journal of Educational Sociology,* 14(6), 358-363.

Fish, L.A. (2013). A comparison of undergraduate versus graduate student perceptions and performance using online homework in Introduction to Operations Management Courses. *Business Education Innovation Journal*, 5(1), 58-66.

Halx, M.D., & Reybold, E. (2005). A pedagogy of force: faculty perspectives of critical thinking capacity in undergraduate students. *The Journal of General Education*, 54(4), 293-315.

Jungic, V., Kent, D., & Menz, P. (2012). On online assignments in a calculus class. *Journal of University Teachings & Learnings*, 9(1), 1-13.

Khanlarian, C.J., & Singh, R. (2014). An exploratory study of the online learning environment. *Issues in Accounting Education*, 29(1), 117-147.

Lockwood, C.A., Ng, P., & Pinto, J. (2007). An interpretive business statistics course encompassing diverse teaching and learning styles. *Academy of Educational Leadership Journal*, 11(1), 11-22.

Maclean, G., & McKeown, P. (2013). Comparing online quizzes and take-home assignments as formative assessments in a 100-level economics course. *New Zealand Economic Papers*, 47(3), 245-256.

Olczak, M. (2014). Evaluating the impact of a web-based resource on student learning. *International Review of Economics Education*, 16, 156-169.

Palocsay, S.W., & Stevens, S.P. (2008). A study of the effectiveness of web-based homework in teaching undergraduate business statistics. *Decision Sciences Journal of Innovative Education*, 6(2), 213-232.

Peters, M., Kethley, B., Bullington, K. (2002). The relationship between homework and Performance in an introductory Operations Management course. *Journal of Education for Business*, 77(6), 340-344.

Piña, Anthony A. (2013). Learning management systems: A look at the big picture. In Y. Kats (Ed.), *Learning Management Systems and Instructional Design: Best Practices in Online Education* (p. 1-19). Hershey, PA: Information Science Reference.

Rhodes, M. Taylor Sarbaum, Jeffrey K. (2015). Online homework management systems: Should we allow multiple attempts? *American Economist* 60(2), 120-131.

Stratling, R. (2015). The complimentary use of audience response systems and online tests to implement repeat testing: A case study. *British Journal of Educational Technology*, 1-15. doi: 10.1111/bjet.12362

Wormeli, R. (2011). Redos and retakes done right. *Educational Leadership*, 69(3), 22-26.

Yourstone, S.A. Kraye, H.S., Albaum, G. (2010). Online quantitative-based tests - are more attempts better for learning? *Decision Sciences Journal of Innovative Education*, 8(2), 347-351.