Journal of the Scholarship of Teaching and Learning, Vol. 16, No. 6, December 2016, pp. 1-18.
doi: 10.14434/josotl.v16i6.20261

Student Achievement in Undergraduate Statistics: The Potential Value of Allowing Failure

Joseph A. Ferrandino[1]

Abstract: This article details what resulted when I re-designed my undergraduate statistics course to allow failure as a learning strategy and focused on achievement rather than performance. A variety of within and between sample t-tests are utilized to determine the impact of unlimited test and quiz opportunities on student learning on both quizzes and subsequent assignments in two undergraduate statistics courses (one a 15-week hybrid and one a 6-week online course). The results show that the previous DFW rate was reduced, and no significant difference in outcomes was found between the two different course modalities. Furthermore, students achieved significantly higher on their last quiz and test attempts on every instrument in both semesters, with large effect sizes. Corresponding assignments showed students achieved significantly higher than the mean first attempt, but significantly lower than final mean quiz attempt scores, suggesting some knowledge was not carried over to application. The article concludes by evaluating the results of unlimited testing with minimum standards and the limitations of the study and the pedagogical model.

Keywords: Learning through Failure, Achievement, Undergraduate Statistics, Online, Minimum Standards

Teaching undergraduate statistics poses many challenges. Twenty five percent of US students that took the ACT in 2012 were rated as not ready for college in all four of the test areas (English, reading, math and science), while 48% were not prepared in reading and 54% were unprepared for college math (ACT, 2012). In this article, I seek to add to a growing body of "teaching statistics" literature after completely overhauling my undergraduate statistics course. First, the literature on teaching statistics is briefly reviewed, followed by the contextual and structural changes made to the course predicated upon allowing failure as a learning tool and a focus on student achievement rather than performance. Next, the data from two courses is discussed, presented and analyzed to determine the impact of the change across two modalities on student completion and learning from a variety of perspectives. The limitations and potential for this approach are discussed in conclusion.

A Review of the "Teaching Statistics" Literature and Its Provocation of Thought

Unfortunately, teaching undergraduate statistics can be difficult for many reasons that have persisted over time, across disciplines (Forte called for an end to "sadistics in statistics" for social

[1] Associate Professor, School of Public and Environmental Affairs (SPEA), Indiana University Northwest, 3400 Broadway, Gary, IN 46408

work students 20 years ago; Forte, 1995) and across borders (see Hindlis and Hronova, 2015 for a discussion from Prague). Several recurrent themes emerge from the literature on teaching basic statistics: motivating students is important, addressing their math anxiety is essential, dealing with performance extremes is expected, overcoming high attrition and failure rates is desired yet difficult, and their exists evidence of limited student ability to grasp difficult concepts and uneven performance across majors (Connors, McCown and Roskos-Ewoldsen, 1998; Chermak and Weiss, 1999; Bushway and Flower, 2002; Proctor, 2006; Forte, 1995; Elliott, Choi and Friedline, 2013, Pan & Tang, 2005).

The pedagogical responses to these problems in the literature are wide and varied. One approach focuses on the learning process relative to enhancing the traditional exchange of information, including strategies such as activity-based learning, mastery-based learning, peer-led team learning, group projects, peer, group and professional tutoring, and learning through technology (Connors, McCown and Roskos-Ewoldsen, 1988; Chermak and Weiss, 1999; Bushway and Flower, 2002; Delucchi, 2007; Curran, Carlson & Celotta, 2013).

Another area of focus is on the utilization of technology to teach statistics in order to improve learning, reduce anxiety and increase student participation. Stickels and Dobbs (2007) found that course structure has an impact on statistics anxiety, as students in a computer-aided statistics class reported significantly lower anxiety than those in a traditional "calculator and paper/pencil" course. There has also been research into specific technological tools and their impact on learning statistics. The use of clickers in several Australian lecture-based undergraduate statistics courses was found to significantly increase student engagement and participation (Dunn, Richardson, McDonald & Oprescu, 2012). Proctor (2002) found that students using Excel exhibited better statistical understanding and increased knowledge than the students randomly assigned to use SPSS. This is important as some evidence suggests more employers prefer Excel, but also seek SPSS, SAS and STATA skills in their new employees (Adams, Lind Infeld, and Wulff, 2013). These findings suggest that the choice of tool taught (Excel, SAS, STATA, Minitab, SPSS, as examples) matters both within and beyond the classroom.

Other professors have introduced changes in their courses to reduce anxiety, especially in testing where it is prevalent (Connors, Mccown and Roskos-Ewoldsen, 1998). Focusing largely on the goal of retention, Bushway and Flower (2002) used multiple choice quizzes with questions (in the form of completed sentences) taken directly from the textbook, but of note was that students could retake quizzes to improve and the questions were randomized, with no question-level feedback provided. This approach is predicated on giving students more opportunity to learn than traditional single-attempt quizzes, while also reducing the anxiety students feel from that common approach.

Finally, the literature also focuses on the most important teaching tool—the professor—in the learning of statistics. Summers, Waigandt and Whitaker (2005) studied the differences in grades and student satisfaction between traditional and online undergraduate statistics courses. They found that there was no difference in student grades, but the online students were significantly less satisfied on several professorial measures (instructor explanations, instructor enthusiasm, instructor openness to students and instructor's interest in student learning) than their in-class counterparts. Other research, using a small sample of focus groups, found that pace of instruction and the instructor's attitude both contribute to statistics anxiety (Pan and Tang, 2005). Furthermore, teaching statistics requires a professor to motivate (Capshew, 2005; Connors, et al, 1998) and engage students (Dunn *et. al*., 2012). Better presentation of information and being

proactive in identifying low and high achievers is important in teaching statistics as well (Connors, Mccown and Roskos-Ewoldsen, 1998).

A review of this literature reveals anxious students, often dreading and/or not completing the course and requiring multiple, evolving approaches to teach them statistics, a subject for which many were unprepared. It also finds professors seeking and testing new ways to reduce anxiety, improve learning and enhance the overall course experience. These issues are largely inter-related, and I found all in my course to varying degrees. I went back and calculated my personal DFW rate (total number of D's, F's and W's out of total students enrolled) for the first four semesters I taught undergraduate statistics: 25.6%. Of all 164 students on the first day roster (average class size was 41 students): 11% had failed for some reason, 8.6% dropped the course before add/drop ended, and 7.9% withdrew sometime after the first week ended for a loss of 42 students. I realized I needed to make wholesale changes in how I utilized technology, delivered information, gave quizzes and tests, encouraged peer interactions and motivated students.

Framework for Redesigning the Course

I started with the concept of the class rather than the content and at the common endpoint of teaching statistics: failure. The great inventor Thomas Alva Edison touted failure as a pathway toward achieving success. Edison is attributed with saying that "many of life's failures are people who did not realize how close they were to success when they gave up" and that he "did not fail, just found 10,000 ways that did not work". Thus, failure can have a positive impact despite its overwhelmingly negative connotation, especially in academia. This is, however, a very natural starting point for many statistics professors and students, especially in light of the literature reviewed earlier that largely focuses on students' failure to learn this subject at high levels after coming in largely unprepared.

Interestingly, the value of failure for its positive impacts is on the rise across a broad spectrum. The *Wall Street Journal* ran an article titled "Better Ideas through Failure" that detailed methods used by several companies to encourage their employees to fail in order to innovate better (Shellenbarger, 2011). A BBC article titled "Secret Google lab 'rewards staff for failure'" discussed how valuable failing is in ultimately succeeding at a company like Google (Grossman, 2014). This should not come as a surprise as Google lists "Fail Well" as its 8th of 9 principles that "any enterprise, large or small, can adopt to steal Google's innovative culture" (Leong, 2013). Author J.K. Rowling of *Harry Potter* fame framed her entire commencement speech to the Harvard class of 2008 around the benefits of failure (Rowling, 2008).

STEM (science, technology, engineering and math) and other educators have also been extolling the learning benefits that failure provide can provide. In his book *Think Like an Engineer*, Mushtak Al-Atabi (2014) promotes encouraging student failure and rewarding those failures to achieve a "return on failure", analogous to a return on an investment. Recently, Kapur (2015) formalized this idea—allowing people to fail as a pathway to better learn something new—and dubbed the approach "productive failure". Oddly, failure is ascendant.

Many students were already failing to learn or complete my course so I had to rethink it. My course would allow and incorporate low stakes failure, for positive means rather than negative ends, while also incorporating higher stakes failure into the course design to motivate students to stay on task, time and in the flow of the course. Failure would be productive for the student as well as the class.

Allowing students to fail as a pathway to success in statistics also requires a concurrent turn away from performance and toward achievement. Performing is defined as "beginning and carrying through to the end…to take action in accordance with the requirements of; to fulfill an obligation or requirement; accomplish something as required or expected" (American Heritage, 2002). By contrast, achievement is defined as "the act of accomplishing or finishing; something accomplished successfully, especially by means of exertion, skill or perseverance" (American Heritage, 2002). I had always given quizzes, tests and assignments and expected performance, i.e. students to show up at a scheduled date and time and "perform", or fulfill one's obligation as a student at the level I expected when required to do so. Thus, performance was measured once on each graded 'artifact' while achievement is an outcome of these successive performances. I had never measured achievement before, only one-time performances.

The paradigm I chose to pursue—incorporating failure and focusing on achievement—is well supported by prior research and practice. I followed the lead of other professors and decided to give every student unlimited opportunities to take quizzes and tests with a minimum score required on each (80% and 70%, respectively) to advance within the course (see Bushway & Flower, 2002; Gunyou, 2015 as examples). This approach is supported by research that espouses the benefits of multiple, spaced, frequent or practice testing as a pathway to increase, improve and/or sustain learning through various mechanisms, methods and conditions (Sparzo, Bennett & Rohm, 1986; Cull, 2000; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b; Karpicke & Roediger, 2007; Karpicke & Roediger, 2008; McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011; Rawson & Dunlosky, 2012). Taken together, these findings point toward using testing as a learning rather than assessment tool and varying the types of tests given, their frequency and their purpose.

Both quizzes and tests were re-created and contained multiple types of questions and random question selection from a created "bank". Any attempt that fell short of the minimum (now a failure), would require the student to re-take the assessment until they achieved a passing grade to continue in the course, theoretically producing the same benefits of practice testing (Rawson & Dunlosky, 2012). Upon meeting at least the minimum standard, students had a choice and could continue taking the assessment until they earned the maximum score or they could choose to move on in the class. Regardless, they took quizzes and tests at their choice of place and time, giving them more control over the process and hopefully reducing anxiety as well. The only caveat was that all attempts had to be completed by the due date and time, and late attempts that exceeded the minimum (8 points) would be scored at the minimum (a late 9 or 10 would revert to an 8 in the gradebook). This approach encouraged students to work ahead (they could move through the course as fast as they wished) but did not allow them to fall too far behind (they could not continue in the course if the minimum was not achieved within a week of the due date for any quiz or test).

This approach to teaching statistics requires a structure that is perfectly tailored to online delivery. I rebuilt the new course in the learning management system using modules of content that were sequentially ordered within each module (see Frantzen, 2014). This method required students to view the lectures (even for the hybrid course) and take quizzes on each new content area (there were thirteen total quizzes worth ten points each). When they reached at least the minimum score, they could advance to a single-submission assignment on the same topic(s). At the end of the last four of the five modules was a module exam worth seventy points each, which could be taken unlimited times until a student scored at least a 49 (70%). This design made sure students were taking the course in order and permitted unlimited opportunity (by the due date for full credit or by the available until date for partial credit) until success was achieved. Students were

not given feedback or incorrect questions on these quizzes or tests until after the period to take it closed, consistent with testing enhancing learning even without feedback provided (Roediger & Karpicke, 2006b). All of the changes of this course were made to reduce student anxiety, measure and improve their learning, provide unlimited opportunity for achievement through allowing failure, and create a better experience overall for more students, consistent with the aims of prior pedagogical research on teaching undergraduate statistics.

The Current Study

The data for this article derives from two undergraduate statistics courses taught in subsequent semesters (spring and summer) using two modalities (15-week hybrid and 6-week online) with identical content and structure. The revised course was implemented during the spring 2015 semester and taught again using the identical course, with the exception that the data used in assignments in summer was changed to prevent student sharing across semesters but the questions remained the same and identical grading rubrics were utilized.

There were 129 students that remained in the courses without withdrawing, and 17 instruments per course with unlimited opportunities to take them. There was a class policy that any late quiz or test above the minimum score would revert back to the minimum as a late penalty. The performance on all quizzes and tests was broken down by class and instrument as well as aggregated by class and instrument to accommodate multiple t-test applications.

Finally, assignment grades were used on each assignment in each class and also aggregated and de-aggregated as needed. All assignments were graded using a rubric worth 30 points: 0-5 points for APA formatting, 0-15 points for the ability to calculate statistics and answer questions properly, and 0-10 points for the ability to use both Excel and SPSS proficiently. The LMS did not permit breaking out these categories, and it was too time consuming and potentially error-inducing to manually enter 1,032 grades for one portion of the rubric individually from the LMS. As with quizzes and tests, late assignments were marked in the gradebook and penalized at 3 points per day, but this was not included in the data export. As a result, the assignment scores (normed 0 – 10) do not represent actual performance but are scored lower to include the late penalties. This makes the performance measure on assignments very conservative here. This data was analyzed using one-sample, paired sample and independent sample t-tests to answer four important questions about to the course revision:

(1) relative to retention and completion, was the DWF rate impacted?
(2) relative to modality differences, were there any discernible differences in outcomes between the two courses?
(3) relative to measuring changes in learning, were the differences in results between first and last quiz and test attempts significant?
(4) relative to the impact of unlimited testing, was there a significant difference between first and last quiz and test attempts and corresponding assignment performance?

Results: The Impact of Unlimited Opportunity on Students and their Learning

Student Completion

One goal of the change in this course was to increase student retention and completion in addition to improved learning and enhanced overall experience. As noted earlier, this course had an established DFW rate of 25.6% (42 of 164 total enrolled students did not complete the course successfully) over the previous four semesters before the course was revised. In the spring hybrid course, the DFW percentage was reduced to 14.6% (7 of 48 enrolled students) and remained 14% (14 of 100) in the summer 6-week fully online course.

This increase in successful completion of the course was a positive indicator of the impact of the course re-design which did not allow students to fall behind or continue if they were not meeting minimum achievement levels. Furthermore, the spring course had six students which had previously been unsuccessful in completing the course. Three of these students (50%) successfully completed the revised course with grades of C or better. Of the other three students, one stopped attending and two completed yet failed due to missed or incomplete work. Both had achieved at least the minimum on all quizzes and tests taken, however one chose not do any of the assignments, which were optional for continuing on in the course. This evidence supports both unlimited opportunity to achieve but inclusion of standards which must be met to avoid failing the course.

Quizzes and Tests

Recall that we are studying two classes so it is critical to test for differences between the two classes to make sure the results are consistent in both modalities of the same course. To accomplish this, the results were aggregated by class, as shown in Table 1 below and compared through a paired sample t-tests of thirteen matched pairs.

For all the first quiz attempts throughout the semester on all 13 quizzes, there was no significant difference between the two courses on mean score (6.1 and 5.9, respectively). There was also no significant difference in aggregate performance on the last quiz attempt (9.24 and 9.01, respectively). The mean difference on each instrument between the first and last attempts was nearly identical and not significantly different (3.15 to 3.10, respectively). There is also no significant difference in either the lower 95% confident interval for the mean difference between first and last attempts on every instrument (2.34 and 2.54, respectively) or the upper 95% confidence interval (3.86 to 3.73, respectively). Finally, the mean effect size indexes for all instruments (.67 and .63, respectively) were not significantly different. Despite the major structural differences in the mode of delivery, the two courses were statistically equivalent on these six measures, initial evidence the changes made to the course are flexible and robust enough overcome differences in how the students take the class.

Table 1: Matched Pairs of Aggregate Class Performance on Quizzes, Between Classes

Measure	Course	Mean (SD)	Mean Difference	SD (SEM)	LL	UL	df	t	d	p
First Quiz Attempts	Summer	5.9 (.75)								
	Spring	6.1 (.90)	-.24	.96 (.27)	-.82	.34	12	-.91	.06	.38
Last Quiz Attempts	Summer	9.01 (.42)								
	Spring	9.24 (.43)	-.19	.40 (.27)	-.43	.05	12	-1.8	.21	.11
Mean Diff (Last – First)	Summer	3.15 (.56)								
	Spring	3.10 (.88)	.05	.97 (.27)	-.54	.63	12	.170	.00	.87

LL 95% CI	Summer	2.57 (.57)								
	Spring	2.34 (.75)	.22	.87 (.24)	-.30	.75	12	.93	.07	.37
UL 95% CI	Summer	3.73 (.55)								
	Spring	3.86 (1.1)	-.12	1.11 (.31)	-.79	.55	12	-.39	.01	.70
Effect Size (d)	Summer	.63 (.11)								
	Spring	.67 (.09)	-.04	.14 (.04)	-.13	.05	12	-1.04	.08	.32

An important initial figure is the number of attempts students made on quizzes and tests when given an unlimited opportunity to take them. Using the traditional testing method, where every student takes each assessment once, this would equate to 2,193 attempts (129 students each taking 17 assessments) if no one missed a single attempt. The previous semester before making this change, students missed 21.2% of all attempts, not including students that withdrew. Instead, this method produced a total of 9,477 attempts. The key to this motivation was in setting minimum performance benchmarks required for advancement in the course that uphold rigor but do not require perfection (80% for quizzes and 70% for tests). In the previous semester before making the change, 178 of the 217 possible attempts students could have taken on the quizzes (82%) were below the newly established 80% threshold. The results also show a high level of students taking advantage of the unlimited opportunities. In the summer course, 81% of test attempts were multiple attempts, ranging from 60.3% of the students on Quiz 12 to 98.4% of the students on Quiz 4.

The initial quiz results were promising along this front as well. Even with the low stakes of unlimited opportunity testing, the percentage of quiz attempts below the threshold to continue in the course was 52.7% (3,552 of 6,742 quiz attempts) and eventually reduced to 0 based on the course structure. By contrast, just 246 attempts (3.6%) were students achieving a perfect score on their first attempt, showing the depth of student need for multiple opportunities. In fact, when given the opportunity to continue working and learning, a total of 789 other quiz attempts (11.7%) eventually reached the perfect score level, with students requiring an average of 4.1 attempts to achieve that result. It also must be noted that initial quiz scores in both revised courses was similar to the initial performance of the class in the previous semester, whose performance averaged between 4.93 and 5.42 on 7 comparable quizzes that could only be taken once. Thus, the three classes all started at the same place, but the latter two classes could move on and achieve learning the former class could not.

Students responded in a variety of ways after reaching the minimum score on each quiz in both semesters. In spring 2015, there were 143 attempts that achieved the minimum quiz score (8 of 10) on the first attempt. Just 17 students stopped there (11.9%) and did not continue. Another 10 students continued but their last attempt was below the minimum and students stopped there (6.9%), which means that just 18.9% of students that had achieved the minimum score on their first attempt chose not to progress past that point before continuing in the course. More encouraging were the 16 times a student continued taking the instrument after initially achieving the minimum score and achieved a 9 of 10 before continuing in the course (11.2%). Finally, 91 students (63.6%), after achieving the minimum on the initial attempt chose to continue until they achieved the maximum score. This means that the ratio of students continuing until they achieved the maximum score to stopping after initially achieving the minimum was just over five to one.

The summer results were similar on this measure. A total of 727 attempts earned at least the minimum score on the first attempt (16.5% of all attempts) and 69% of the time students took at least one more attempt. There were 225 occasions where a student earned at least the minimum score on their initial attempt and did not take another attempt (5% of all attempts). Of these, 53 students earned an 8 and stopped there; 16 earned a 9 and stopped there and 156 scored a 10 (maximum score) and did not need to take it again. However, 502 times a student earned at least the minimum and took at least one more attempt. These results show that students will overachieve their initial results if given the chance, though some will stop before reaching that point. However, all are required to at least reach that minimum. Further, it shows that despite low stakes testing and failure as an option, many students do perform well on the initial attempt and those that do not are provided an opportunity to overcome the initial failure and even achieve perfect scores should they decide to do so.

Finally, for each assessment (N = 17), every student that had more than one attempt on the quiz or test was included in their respective class' sample, and a paired sample t-test was utilized to determine the difference in performance between the first and last attempt. As shown in Table 2 below, student performance significantly increased on every instrument in both courses, with the average effect size .67 in the spring and .63 in the summer. Most important was the fact that these courses were taught using two seemingly different formats but achieved similar results. These findings provide evidence that learning improved from this approach in both quantity (attempts) and quality (significant increases in comparative performance at two times). Furthermore, though not long term, these results indicate that in the short term (initial to final attempts) students scored significantly higher, supporting this type of testing to improve rather than evaluate learning.

The module tests also serve an important function in the course, but their discussion here is limited in scope. As shown in Table 2 below, test results mirrored quiz results as students showed significantly higher scores on their last test attempt than their first for each test in each course, with large effect sizes ranging from .42 to .77 and mean differences ranging from +9.4 to +20.8. In relation to failure and achievement, the final test was the only one in either course to not have a minimum, meaning students were only required to submit, not achieve, with failure an option as an end in itself. Between the two courses, achievement dropped considerably, as did the initial test results. This finding suggests that without the minimum standard of achievement in place, failure as a learning strategy was not effective. It appears students, knowing they did not have to achieve at a specified level, did not take initial attempts seriously. Both classes had extremely low mean attempt scores and 42 of the 114 students with more than one attempt on the test (36.8%) chose an achievement level lower than the 70% threshold of the previous exams. While students took many attempts on the exam (862 of 2,742 total test attempts, or 31.4% of all test attempts), just 16% of attempts scored at or above the minimum 70% threshold, the lowest of any test. Thus, this particular test showed that without the minimum achievement requirements, many students will opt for a failing grade rather than using failure as part of the learning process.

Assignments

While the tests and quizzes were mandatory and low stakes by design, the assignments were high stakes with only a single submission allowed per student (N = 8) and a late penalty equal to 3 points per day off the earned grade. This method permitted students to be fully prepared for the assignments after having watched (and/or heard in person and watched) the lectures, passed the corresponding quiz or set of quizzes with at least an 80%, watched a tutorial on how to complete

the assignment and participate in the discussion board as they worked. Thus, assignments were a clear yet short term measure of aggregate student learning and could be linked with the lectures and quizzes that were designed to help students learn statistical language, concepts and tools and apply them afterward.

As with the quiz and test performance, this analysis began by comparing the assignment grades between the two classes. Each class used the identical grading rubric and the assignments were identically worded but each class used entirely different data to prevent students from sharing results from the previous course. As shown in Table 3 below, 6 of the 8 assignments showed no significant difference between the mean scores using an independent sample t-test. For Assignment 1, students in the summer session scored significantly higher than in the spring (8.71 to 8.17, respectively) though the spring students scored significantly higher on Assignment 7 than their summer counterparts (8.11 to 7.33). These results were surprising given the spring class met in person once per week and had nine more total weeks of class time. This provides further evidence that students performed equally across modalities, with some slight variation that will always be present in any classroom.

For both classes, there were 8 total assignments and 129 students, for a possible total of 1,032 assignments. Of these, just 65 were not submitted at all (6.3%), with some of these being missed because the student stopped attending and/or participating, prohibiting them from advancing in the course. An additional 163 assignments (15.8%) were submitted late. In all, this initial submission data was very positive as the assignments were not required to continue in the course and left entirely up to the student to complete and submit, with 93.7% of assignments submitted and 78% submitted on time.

Table 2: Paired Sample T-Test Results for all Quizzes and Tests, By Student with 2+ Attempts, Within Courses

Test	Course	First Attempt Mean (SD)	Last Attempt Mean (SD)	Mean Diff	SD	t	df	p	95% Confidence Interval		Effect Size[a]	Mean/Med # Attempts	Min–Max Attempts
Quiz 1	Spring	7.80 (1.2)	9.90 (0.35)	+2.10	1.07	11.56	36	<.001***	1.67	2.38	0.79	3.24/3	1 - 7
	Summer	6.28 (2.0)	8.81 (1.78)	+2.53	2.49	8.261	65	<.001***	1.91	3.14	0.51		
Quiz 2	Spring	5.58 (2.0)	8.88 (0.94)	+3.30	2.16	9.70	39	<.001***	2.61	3.98	0.71	9.56/7	1 - 25
	Summer	5.95 (2.4)	8.85 (1.37)	+2.89	2.84	8.99	77	<.001***	2.25	3.53	0.51		
Quiz 3	Spring	6.31 (1.5)	8.54 (1.30)	+2.30	1.85	7.13	34	<.001***	1.60	2.87	0.60	8.56/6	1 - 18
	Summer	5.13 (1.9)	8.54 (1.27)	+3.41	2.03	14.67	76	<.001***	2.94	3.87	0.74		
Quiz 4	Spring	5.92 (1.4)	8.36 (0.99)	+2.40	1.71	8.88	38	<.001***	1.88	2.99	0.67	15.37/12	1 - 33
	Summer	4.12 (2.6)	8.61 (0.95)	+4.50	2.59	16.19	86	<.001***	3.94	5.05	0.75		
Quiz 5	Spring	7.46 (1.5)	9.44 (1.10)	+1.98	2.01	6.11	38	<.001***	1.32	2.63	0.50	5.38/5	1 - 9
	Summer	6.18 (2.3)	9.20 (1.05)	+3.01	2.50	10.36	72	<.001***	2.43	3.60	0.60		
Quiz 6	Spring	6.95 (1.4)	9.05 (1.66)	+2.11	2.08	6.25	37	<.001***	1.42	2.79	0.51	4.76/4	1 - 9
	Summer	6.63 (1.8)	8.88 (2.18)	+2.25	2.70	6.70	64	<.001***	1.58	2.92	0.41		
Quiz 7	Spring	5.59 (2.5)	9.17 (1.14)	+3.59	2.10	9.21	28	<.001***	2.79	4.39	0.75	5.66/5	1 - 16
	Summer	5.03 (2.3)	8.57 (2.01)	+3.53	2.79	9.813	59	<.001***	2.81	4.25	0.62		
Quiz 8	Spring	4.88 (2.7)	9.38 (1.07)	+4.50	2.38	10.69	31	<.001***	3.64	5.36	0.79	3.62/3	1 - 33
	Summer	5.58 (2.5)	8.80 (2.12)	+3.22	2.63	10.01	66	<.001***	2.59	3.86	0.60		
Quiz 9	Spring	4.70 (2.3)	9.14 (1.46)	+4.43	2.63	10.25	36	<.001***	3.56	5.31	0.74	8.42/6	1 - 27
	Summer	5.29 (2.4)	8.71 (2.27)	+3.41	2.99	10.16	78	<.001***	2.75	4.09	0.57		
Quiz 10	Spring	5.88 (1.9)	9.21 (1.57)	+3.33	2.32	8.23	32	<.001***	2.50	4.20	0.68	3.24/3	1 - 13
	Summer	6.38 (2.0)	9.35 (0.86)	+2.97	1.92	11.96	59	<.001***	2.47	3.46	0.71		
Quiz 11	Spring	6.35 (2.5)	9.56 (0.79)	+3.20	2.24	8.36	33	<.001***	2.42	3.99	0.68	4.66/4	1 - 10
	Summer	6.54 (2.0)	9.41 (1.20)	+2.87	2.00	10.65	54	<.001***	2.33	3.41	0.68		
Quiz 12	Spring	6.27 (2.3)	9.48 (1.42)	+3.21	2.23	7.48	26	<.001***	2.32	4.09	0.68	4.63/2	1 - 69
	Summer	6.21 (2.0)	9.70 (0.71)	+3.49	1.97	12.65	50	<.001***	2.93	4.04	0.76		
Quiz 13	Spring	5.62 (2.7)	9.55 (1.33)	+3.93	3.14	6.74	28	<.001***	2.73	5.13	0.62	3.59/3	1 - 22
	Summer	6.86 (1.5)	9.75 (0.90)	+2.89	1.80	12.99	64	<.001***	2.45	3.33	0.73		
Test 1	Spring	48.03 (6.26)	57.44 (6.26)	+9.40	7.40	8.04	39	<.001***	7.03	11.77	0.62	6.75/7	1 - 12
	Summer	44.47 (14.8)	56.21 (13.8)	+11.75	13.90	7.84	85	<.001***	8.77	14.73	0.42		
Test 2	Spring	49.23 (9.14)	58.67 (11.1)	+9.40	10.52	5.30	34	<.001***	5.81	13.04	0.45	6.52/5	1 - 19
	Summer	37.87 (14.6)	52.16 (6.20)	+14.29	13.31	9.46	69	<.001***	11.15	17.45	0.56		
Test 3	Spring	49.59 (9.30)	65.08 (6.79)	+15.50	8.49	11.10	36	<.001***	12.66	18.32	0.77	4.86/4	1 - 16
	Summer	51.67 (11.0)	63.53 (7.93)	+11.86	9.46	9.71	59	<.001***	9.42	14.31	0.62		
Test 4	**Spring**	**28.5 (12.07)**	**49.32 (11.7)**	**+20.84**	**13.21**	**9.85**	**38**	**<.001*** **	**16.56**	**25.12**	**0.72**	**8.66/5**	**1 - 70**
	Summer	**31.23 (14.1)**	**48.86 (12.9)**	**+17.62**	**15.42**	**9.96**	**75**	**<.001*** **	**14.10**	**21.14**	**0.57**		

[a] d is a measure of effect size, calculated as $t^2 / (t^2 + df)$

Table 3: Independent Sample T-Tests of Assignments, Between Courses

Assignment	t	df	Mean Difference	Std. Error Difference	95 % CI LL	95 % CI UL
1[a]	-2.298*	51.235	-.53968	.23487	-1.01	-.068
2	.843	127	.21217	.25158	-.286	.710
3	.229	121	.05691	.24864	-.435	.549
4	-.112	122	-.04413	.39352	-.823	.734
5	-.386	109	-.17381	.45028	-1.06	.718
6	.987	117	.25541	.25870	-.256	.767
7	2.596**	110	.79089	.30467	.187	1.39
8	1.612	121	.37331	.23165	-.085	.831

a denotes that this was the only test where equal variances were not found, so adjusted results are presented
**p < .01; *p < .05

Table 4: One Sample T-Test Comparing Assignment Performance to First and Last Quiz Attempt Benchmarks, within Courses

Assignment	Corresponding Quiz	Course	Mean Assignment Grade	Mean First Quiz Attempt	Mean Last Quiz Attempt
1	1	Spring	8.17	8.0	9.88***
		Summer	8.71	6.29***	8.80
2	2	Spring	8.66	5.58***	8.88
		Summer	8.44	5.89***	8.90**
3	3,4	Spring	7.94	6.10***	8.45*
		Summer	7.89	4.49***	8.52***
4	5,6,7	Spring	7.81	6.76***	9.23***
		Summer	7.85	6.09***	8.84***
5	8,9	Spring	6.74	4.78***	9.24***
		Summer	6.92	5.33***	8.87***
6	10	Spring	7.71	5.88***	9.21***
		Summer	7.46	6.81***	9.43***
7	11	Spring	8.11	6.35***	9.56***
		Summer	7.33	6.35***	9.46***
8	12,13	Spring	7.25	5.94***	9.51***
		Summer	6.88	6.50*	9.75***

***p < .001; **p < .01; *p < .05

To determine the impact that unlimited quizzes had on student performance on assignments, a series of one sample t-tests were utilized. The course was revised to require students achieve at least an 8 of 10 (80%) on each concept quiz before moving on to the assignment that would assess their ability to apply that conceptual knowledge (no more than three concepts per assignment). Each assignment corresponded with specific quizzes, as detailed in Table 4 below. For each course, student performance on each assignment was compared to the class mean for the first quiz attempt (initial knowledge) and last quiz attempt (final knowledge), which represented the population parameter used to test each sample against on each assignment.

These tests reveal some interesting results. For the spring class, students significantly outperformed their initial quiz scores on each assignment except the first. This shows that students were able to apply the information learned at a higher level than their initial conceptual knowledge would indicate. On average, students outperformed their initial quiz attempts on the same information by 1.62 points out of 10 (mean assignment score of 7.79 to mean first quiz attempt score of 6.17). However, there appeared to also be a significant amount of knowledge not carried over from the conceptual to the application stages of learning, perhaps as expected. On 7 of the 8 assignments, the mean assignment score was significantly lower than the mean final quiz score achieved by the students. The exception was Assignment 2 in which students scored significantly higher than the first quiz attempt mean (MD = +3.08) but statistically equal to the mean last quiz attempt (MD = -.22). In all, the average assignment score was -1.44 points lower than the mean final quiz score. Thus, assignment score was almost directly between the mean first quiz score (initial assessment of understanding) and mean last quiz attempt score (final assessment of understanding) yet statistically different from both. This finding perhaps represents a true regression to the mean in terms of student learning, where application is significantly above initial knowledge, and significantly below final knowledge level.

The results from the summer class were very similar despite the difference in course delivery. These students also outperformed the mean initial quiz attempt score by an average of 1.7 points on corresponding assignments (mean assignment score is 7.68, mean first quiz score of 5.98). This difference was significant on every single assignment, though the last assignment in summer was very close to statistical equivalence. In regards to the mean last quiz attempt, the summer class performed statistically equal to on the first assignment (8.8 to 8.7, respectively). However, on all other assignments, the class score was significantly lower than the mean last quiz attempt score, with a mean difference of -1.39 (mean assignment score of 7.68 compared to mean last quiz attempt score of 9.07). As with the spring class, this shows that assignment scores were significantly higher than initial mean quiz attempts but significantly lower than mean last quiz attempts. The quizzes had a significant impact, but a certain and significant percentage of conceptual knowledge was not transferred to the application stage as measured by assignment performance. The caveat here is that late penalties were included, as were all three parts of the grading rubric, meaning final assignment scores were lower and quite conservative here.

Limitations

Though nearly all research into this topic use classes taught by the authors, this represents a limitation that must be addressed. As a result of this approach, results may not be generalizable to other campus contexts or learning management systems.

There are also data issues to discuss. The learning management system used is not structured to export the type of data I needed to analyze, so much of it had to be created. This

limits the type of information that can be gleaned from the data without extensive filtering and rechecking as was the data structured for this analysis. For example, late penalties could not be separated easily from the final score on quizzes, tests and/or assignments, especially in such large classes. This limitation is mitigated by the fact that this would deflate grades, not inflate them, making any significant findings more valid, especially when compared to final attempts. Another example is the fact that the attempt data for quizzes and tests were separated by instrument and had to be entirely reformatted and enhanced to conduct the analyses needed here.

Finally, there is always the limitation of the difficulty or quality of the instruments, which is rarely discussed in the literature. This type of course structure has the added benefit of allowing much more difficult questions than the traditional way of testing, and this is even more encouraged when students are allowed to fail as a pathway to success. As a note, the quizzes, tests and assignments were all more difficult than previous versions of the course, but there is no way to quantify or compare this statement to other classes studied in the literature.

Finally, this teaching approach would not work in a traditional classroom setting as testing would take up nearly all the class time available. This is a limitation as applying this method in face-to face teaching is just simply not practical given the time constraints. In addition, the LMS did not allow for a mandatory "cooling off period" between attempts that would force students to take time between quiz or test attempts. This is unfortunate as spacing tests is known to produce learning benefits (Cull, 2000). Furthermore, learning was not measured after the course ended to determine information retention. This all must be considered relative to the results discussed and the outcomes assessed.

Discussion and Implications for the Teaching and Learning of Statistics

The results found in my two courses validated my decision to overhaul my statistics course and entirely re-design it from a perspective of failure and achievement. As student retention and learning outcomes are administrative focal points and statistics typically has a high DFW rate that impacts the learning process, I was pleased to reduce my personally unacceptable DFW rate almost in half in both courses. This means more students completing, and by default more students taking part in the learning process. I was equally pleased as this was accomplished while simultaneously making the course content more difficult and rigorous, showing that one need not be sacrificed for the other. This measure can also provide a proxy for anxiety levels as it seems valid on its face that higher anxiety would lead to more DFW within the course. In addition, students had greater control over their grade and learning, and anecdotally many students pointed to the class structure as contributing to their retention and success in the course, specifically the discussion boards and the reduced anxiety relative to tests and quizzes.

Equally important was the finding that two classes with entirely different modalities used the exact same course structure and content and achieved nearly identical and statistically equivalent results on almost all measures. This validates the philosophical foundations of the course revision relative to encouraged failure and forced achievement. A pedagogical approach that transcends mode of delivery is both efficient and equitable and provides broader avenues for student success in either format. That said, it is highly unlikely this method could be transferred to traditional in-person teaching without some online component.

The sheer number of quiz and test attempts (9,477) showed a high level of student determination to achieve, but many students unfortunately were content to get the minimum, or slightly above, and not continue despite having a chance to earn a 100% on every instrument that

comprised roughly 66% of the course grade. As the average mean quiz score difference between the first and last attempt was +3.1, it took an additional 5,065 total attempts (or an average of 389 per instrument) to achieve that difference, evidence of an incremental increase in learning that singular high stakes testing does not permit. Though learning was significantly improved between each instrument on the first and last attempt, a percentage of this learning was not fully transferred in the transition to the application in assignments, though some was.

In addition, the results from the last test, the only one without a minimum score requirement, showed that students will fail to achieve if not given benchmarks or standards to achieve to. In the spring semester for example, 12 of 38 students with more than one attempt (31.6%) scored below the minimum 70% threshold and stopped taking the test. Paradoxically, in both classes the last test had the most attempts with the least overall achievement. This finding supports the decision to enact minimum performance benchmarks to encourage achievement by learning through failure rather than allowing students to fail, many of whom will without the minimum requirement standards. Thus, the structure of the course is important in statistics and allowing failure on its own may not produce learning if not combined with a focus on achievement, i.e. meeting minimum standards along the way.

That said, pre-tests in both classes that incorporated two random questions from each of the five modules found initial student knowledge in the 4.4 range (see Figure 1 below), well below the mean initial quiz attempt score in the semester (6.0). This suggests that the lectures and course material increased initial student performance by about 1.6 points on average. Unlimited testing, as measured by the mean last quiz attempt score, increased pre-test scores by +4.8 points and initial topical learning by about +3.1 points. However, this process increased knowledge as measured by mean assignment scores, by +3.3 points from pre-test levels (4.4 to 7.7, respectively), +1.7 points on initial quiz attempts and -1.9 points from final mean quiz attempts. That gaining of knowledge suggests learning improved, but the loss of knowledge from the conceptual to application stages means that the unlimited attempts on quizzes may have a peak return per attempt, after which attempts are made that do not translate into learning gains on assignments. Interestingly, students were surveyed before the class started and asked to predict their final grades, and the class average across both courses was 8.5 of 10, meaning pre-test and initial topical quiz performance were well below student expectations of performance while final quiz attempts exceeded this prediction benchmark. As shown in Figure 1, students were able to achieve their prediction levels in the aggregate despite starting both the course and the material at fairly low performance levels. High anxiety levels are undoubtedly found when student expectation of performance is not possible to achieve based solely on performance levels like those found on first quiz attempts, and failure is an outcome rather than part of the learning process. That said, effort diminishes greatly when the minimum standards are not enforced. In addition, as found in previous research, actual initial student performance was not correlated with student's predictions of their performance (see Karpicke & Roediger, 2008), showing the need for a focus on achievement to better align the two measures.

There are two major drawbacks to this method that perhaps impact the results found here. The first is the structure of the course. Mandatory scores on quizzes were required to advance to the assignments (which were not mandatory to continue), and would also open the exam that required a 70% to move to the next module. This decision was made to give students more time on the assignments as they were high stakes and very in-depth, but it could be more beneficial to have them take the exams first to increase learning for the assignments. An attempt was made to not over-regulate the course and give students some choice in how they approached the

assignments and exams. As shown, initial test attempts also show a significant decrease from final quiz attempts on the same content areas, meaning there may need to be some adjustment to the exam process to make the quizzes more meaningful (for example, three attempts with the last attempt going in the gradebook rather than a minimum score and unlimited opportunities).

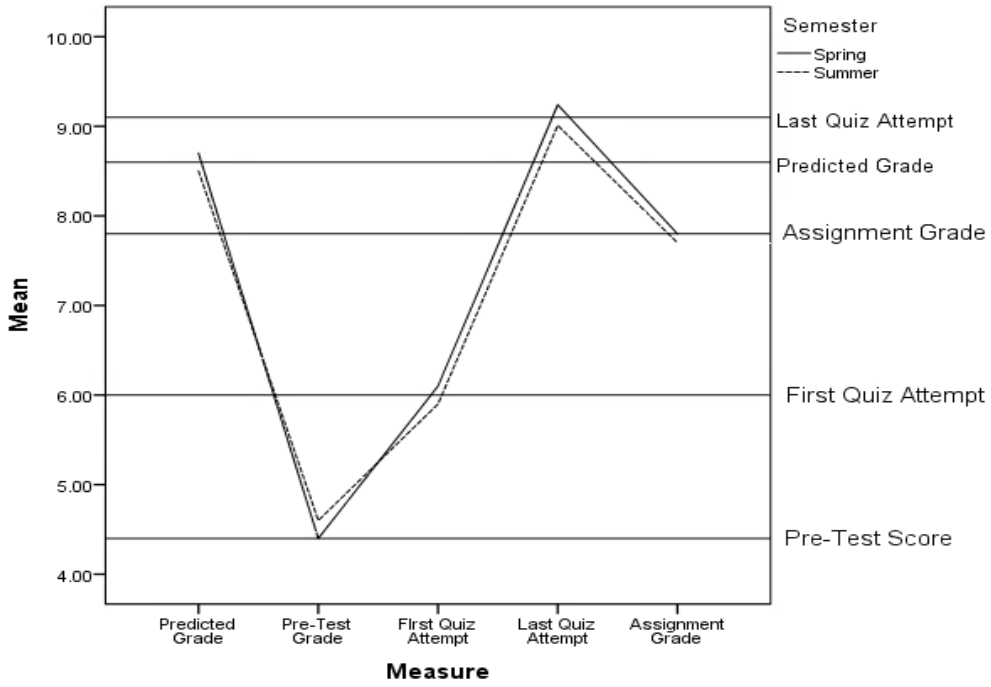

Figure 1: Aggregate Student Grade Prediction, Pre-Test Scores, Initial and Final Quiz Attempts and Assignment Outcomes

Conclusion

The results of this study, though limited by the data and the focus on two classes at one university, are in line with other studies of undergraduate statistics and provide support for a continued inclusion of failure as learning pathway and a focus on achievement rather than performance. As with many other studies, this course exposed areas of weakness in this pedagogical approach that will need to be addressed as the course and I evolve together. That said, the overall experience was far more positive than the previous version of the course, the results were equivalent in two different modalities, more students stayed in the course and completed successfully and the results of this course revision support its undertaking.

References

ACT (2012). *The Condition of College and Career Readiness National Report.* Found online at http://www.act.org/newsroom/data/2012/pdf/CCCR12-NationalReadinessRpt.pdf.

Adams, W.C., D.L. Infeld & C.M. Wulff (2013). Statistical software for curriculum and careers. *Journal of Public Affairs Education.* 19(1), 173-188.

American Heritage College Dictionary (2002). 4th Ed. Houghton Mifflin Harcourt: Boston, MA.

Al-Atabi, M. (2014). *Think like an engineer*: *Use systematic thinking to solve everyday challenges & unlock the inherent values in them.* Creative Commons. As cited by Mushtak Al-Atabi in a blog post titled "Return on Failure: Why I encourage my students to fail (and reward them for that)" on September 26, 2014 found online at: https://www.linkedin.com/pulse/20140926152625-39413468-return-on-failure-why-i-encourage-my-students-to-fail.

Bushway, S.D. & S.M. Flower (2002). Helping criminal justice students learn statistics: A quasi-experimental evaluation of learning assistance. *Journal of Criminal Justice Education.* 13(1), 35-56.

Capshew, T.F. (2005). Motivating social work students in statistics courses. *Social Work Education.* 24(8), 857-868.

Chermak, S. & A. Weiss (1999). Activity-based learning of statistics: Using practical applications to improve students' learning. *Journal of Criminal Justice Education.* 10(2), 361-372.

Connors, F.A., S.M. Mccown & B. Roskos-Ewoldsen (1998). Unique challenges in teaching undergraduate statistics. *Teaching of Psychology.* 25(1), 40-42.

Cull, W.L. (2000). Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall. *Applied Cognitive Psychology.* 14, 215-235.

Curran, E., K. Carlson & D.T. Celotta (2013). Changing attitudes and facilitating understanding in the undergraduate statistics classroom: A collaborative learning approach. *Journal of the Scholarship of Teaching and Learning.* 13(2), 49-71.

Delucchi, M. (2007). Assessing the impact of group projects on examination performance in social statistics. *Teaching in Higher Education.* 12(4), 447-460.

Dunn, P.K., A. Richardson, C. McDonald & F. Oprescu (2012). Instructor perceptions of using mobile-phone-based free classroom response system in first-year statistics undergraduate courses. *International Journal of Mathematical Education in Science & Technology.* 43(8), 1041-1056.

Elliott, W., E. Choi & T. Friedline (2013). Online statistics labs in MSW research methods courses: Reducing reluctance toward statistics. *Journal of Social Work Education.* 49(1), 81-95.

Forte, J.A. (1995). Teaching statistics without sadistics. *Journal of Social Work Education.* 31(2), 204-218.

Frantzen, D. (2014). Is technology a one-size-fits-all solution to improving student performance? A comparison of online, hybrid, and face-to-face courses. *Journal of Public Affairs Education.* 20(4), 565-578.

Grossman, D. (2014). *Secret Google lab 'rewards staff for failure*. BBC News (January 24, 2014). Found online at http://www.bbc.com/news/technology-25880738.

Gunyou, J. (2015). I flipped my classroom: One teacher's quest to remain relevant. *Journal of Public Affairs Education.* 21(1), 13-24.

Hindlis, R. & S. Hronova (2015). Are we able to pass the mission of statistics to students? *Teaching Statistics.* 37(2), 61-65.

Karpicke, J.D. & H.L. Roediger (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language.* 57, 151-162.

Karpicke, J.D. & H.L. Roediger (2008). The Critical Importance of Retrieval for Learning. *Science.* 319, 966-968.

Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice.* 1(1), 58-65.

Leong, K.C. (2013). "Google Reveals its 9 Principles of Innovation." (November 20, 2013). Available online at https://www.fastcompany.com/3021956/how-to-be-a-success-at-everything/googles-nine-principles-of-innovation.

McDaniel, M.A., P.K. Agarwal, B.J. Huelser, K.B. McDermott & H.L. Roediger (2011). Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement. *Journal of Educational Psychology.* 103, 399-414.

Pan, W & M. Tang (2005). Students' perceptions on factors of statistics anxiety and instructional strategies. *Journal of Instructional Psychology.* 32(3), 205-214.

Proctor, J.L. (2002). SPSS vs. Excel: Computing software, criminal justice students and statistics. *Journal of Criminal Justice Education.* 13(2), 433-442.

Proctor, J.L. (2006). Academic achievement and statistical knowledge: A comparison of criminal justice and noncriminal justice majors. *Journal of Criminal Justice Education.* 17(1), 143-161.

Rawson, K.A. & J. Dunlosky (2012). When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning? *Educational Psychology Review.* 24, 419-435.

Roediger, H.L. & J.D. Karpicke (2006a). Test-Enhanced Learning: Taking Memory Tests

Improves Long-Term Retention. *Psychological Science.* 17(3), 249-255.

Roediger, H.L. & J.D. Karpicke (2006b). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science.* 1(3), 181-210.

Rowling, J.K. (2008). Commencement Speech to the Harvard Class of 2008. Published by the Harvard Gazette on June 5, 2008 and found online at http://news.harvard.edu/gazette/story/2008/06/text-of-j-k-rowling-speech/

Shellenbarger, S. (2011). "Better ideas through failure: Companies reward employee mistakes to spur innovation, get back their edge". *The Wall Street Journal*, September 27, 2011. Found online at http://www.wsj.com/articles/SB10001424052970204010604576594671572584158

Sparzo, F.J., C.M. Bennett & R.A. Rohm (1986). College student performance under repeated testing and cumulative testing conditions: Report on five studies. *The Journal of Educational Research.* 80(2), 99-104.

Stickels, J.W. & R.R. Dobbs (2007). Helping Alleviate Statistical Anxiety with Computer Aided Statistical Classes. *Journal of the Scholarship of Teaching and Learning.* 7(1), 1-15.

Summers, J.J., A. Waigandt & T.A. Whittaker (2005). A comparison of student achievement and satisfaction in an online versus a traditional face-to-face statistics class. *Innovative Higher Education.* 29(3), 233-250.