9-13-2019

If at First You Do Not Succeed: Student Behavior When Provided Feedforward With Multiple Trials for Online Summative Assessments

Emily Faulconer
Embry-Riddle Aeronautical University, faulcone@erau.edu

J. C. Griffith
Embry-Riddle Aeronautical University, griff2ec@erau.edu

H. Frank
Embry-Riddle Aeronautical University, frankh1@my.erau.edu

Scholarly Commons Citation

Faulconer, E., Griffith, J. C., & Frank, H. (2019). If at First You Do Not Succeed: Student Behavior When Provided Feedforward With Multiple Trials for Online Summative Assessments. *Teaching in Higher Education, 26*(4). https://doi.org/10.1080/13562517.2019.1664454

If at first you do not succeed: student behavior when provided feedforward with multiple trials for online summative assessments

Emily Faulconer[a]*, J.C. Griffith[a], H. Frank[a]

[a]*Math, Physical, and Life Sciences Department, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA*

* 600 S. Clyde Morris Boulevard, Daytona Beach, FL 32114, U.S.A.; faulcone@erau.edu

Biographical Notes

Dr. Emily Faulconer is the chair of the physical and life sciences discipline within the Math, Physical, and Life Sciences department at Embry-Riddle Aeronautical University. She earned a Ph.D. in Environmental Engineering Sciences from the University of Florida in 2012. Her research interests are within the Scholarship of Teaching and Learning, primarily in undergraduate research and online education. Safety is also an area of interest, and she serves as the Chair of the Academic Safety Committee for Embry-Riddle Aeronautical University. She is also actively involved in national level service through the American Chemical Society’s Science Coach Program and the Journal of College Science Teaching’s Advisory Board.

Dr. John Griffith is the Department Chair for the Mathematics and Physical Life Sciences Department at Embry-Riddle Aeronautical University – Worldwide Campus. He earned his Ph.D. from the University of North Texas and published several articles on distance learning. Dr. Griffith has won research awards for his work in the areas of teaching study skills to students and comparing student and teacher perceptions regarding distance learning.

Hayden Frank is a student at Embry-Riddle Aeronautical University - Worldwide Campus. He is pursuing a Bachelor's of Science in Aeronautics and teaches pilots and aircrew in a flight simulator in Fort Worth, Texas. Hayden has over five years of experience in curriculum development and teaching with emphasis on evaluation analysis and revision.

If at first you do not succeed: student behavior when provided feedforward with multiple trials for online summative assessments

Best practices suggest that timely, actionable feedback is provided with the option to apply the feedback. We used a learning management system to deliver assessments with automatic feedback provided at the conclusion of the assessment, allowing for multiple attempts in order to apply the knowledge gained. Questions were pooled so each attempt was unique, the highest score earned was awarded, with no penalty for failure to use multiple attempts. We found that students who did not earn an A on their first attempt were more likely to try again. Those that did tended to score better on their second attempt. This leads us to conclude that assessment design with multiple attempts that incorporates feedforward influences student behavior. Future work will include additional STEM general education courses in a broader study and a survey of student opinions regarding the utility of the feedback and the option for multiple attempts.

Keywords: multiple attempts, summative assessment, feedback, feedforward

Introduction

Online assessments can be formative or summative, depending upon their use in the course (Sewell, Frith, & Colvin, 2010). Summative assessment occurs at the end of the learning process and evaluates students' mastery to determine a grade. Formative assessment is a part of the learning process, used as a diagnostic tool to provide feedback on progress towards mastery of learning objectives. Learning management systems (LMS) are commonly used as a platform for summative assessments (Coates, James, & Baldwin, 2005; Stodberg, 2012). Instructors can customize the design and deployment of the assessment by selecting how questions are presented (one at a time versus all at once), pulling from question pools, setting assessment availability (synchronous versus asynchronous access), and programming grade availability (immediately after completion or after instructor review). Exemplary assessments are engaging and guide the students in the learning process, with the following characteristics: 1) valid, 2) coherent, 3) authentic, 4) rigorous, 5) engaging, 6) challenging, 7) respectful, and 8) responsive. The design of e-assessments must consider underlying pedagogy (Huba & Freed, 1999).

One way to turn an assessment into an engaging and guiding learning opportunity is to program feedback into the LMS for timely dissemination. According to a 2014 study, students perceived automatically generated feedback as substantially more constructive than manual feedback (Bayerlein, 2014). While timely feedback is a demonstrated best practice (Gaytan & McEwen, 2007; G. P. Wiggins, 1993), the rapidness of the automatically generated feedback was not perceived as a significant benefit by students (Bayerlein, 2014). As opposed to manual feedback which is inherently varied, automatic feedback can be consistently phrased in supportive language, closely aligned with assessment criteria, aimed at the gap between instructor expectations and student performance, and focused on specific recommendations, all of which contribute to the engaging nature of the feedback (Bayerlein, 2014). Linking feedback to the learning process and providing opportunity for application of the feedback is ideal (Hughes, 2011; Little, Bjork, Bjork, & Angello, 2012; G. Wiggins, 2012).

While much emphasis in the literature has been placed on feedback, there is increasing focus on the use of feedforward. This pedagogical technique provides an opportunity for timely application of feedback. In feedforward, students interpret and apply the instructor’s feedback in order to close the performance gap and to improve their demonstration of mastery of learning objectives (Dulama & Ilovan, 2016; Goldsmith, 2008; Koen, Bitzer, & Beets, 2012; Rodriguez-Gomez & Ibarra-Saiz, 2015). Feedforward also serves an important role in clarifying instructor expectations (Baker & Zuvela, 2013). This technique establishes a learner-centered environment that stimulates active learning.

Assessments can themselves be an empowering learning tool. The practice of providing multiple attempts on assessments is under-explored in the literature. Several peer-reviewed studies explored this topic, all with varied parameters (testing time and access, provision of feedback, and scoring of multiple attempts.). Only 36.5% of students completed a second attempt for online homework in an operations management course and they did not outperform one attempt (Orchard, 2016). Effectiveness of multiple attempts in this study were likely limited because feedback was only visible after final submission and the final grade was based on the final attempt, not the higher score. In a macroeconomics course, student scores on homework and exams improved with a second attempt, with similar or slightly higher time on task (Rhodes & Sarbaum, 2015). An indication of missed topics was provided after the first attempt but detailed feedback was provided after final submission. This study also used identical questions on each attempt rather than pools. Students earned the higher of the two grades, though the authors noted evidence of reduced effort in order to make use of information provided on the 1st attempt to improve grades in subsequent attempts, leading them to suggest the averaging of scores to remove the incentive for this less desirable behavior. Students completing two attempts for online homework, quizzes, and exams in an introductory operations management course outperformed those who were allowed four attempts (Yourstone, Kraye, & Albaum, 2010). The use of feedback was not expressly discussed by authors but personal communication with authors revealed it was provided within the LMS. Similar to Rhodes & Sarbaum (2015), this study also observed experimentation by students with a throwaway attempt. A newsletter communication reported utilization of a second attempt on assessments ranging from 35 – 65%, with students keeping the highest score (Luebben, 2008). Average gains ranged from 0 – 10%, though no feedback was provided between attempts. Gamesmanship with a throwaway attempt was also witnessed.

Studies exploring multiple attempts on certification, licensure, or placement exams presented results relevant to this study. A simulation study found that the expected pass rate increases with more attempts (testing volume defined as total number of examinees) (Cheng & Cheng, 2016). A study of credentialing exams found that examinee ability remained consistent when retesting with an identical assessment versus a parallel one (Feinberg, Raymond, & Haist, 2015). Interestingly, this study found that examinees selected the same wrong response on the second attempt 68% of the time, implicating the need for feedback and remediation in the retesting process. A study of nursing school entrance exams found that scores on repeat attempts increased significantly regardless of whether the same or a parallel exam was provided (Wolkowitz, 2011). Additionally, no influence of lag was found, with examinees performing similarly regardless of the number of days between attempts.

The literature is incredibly scarce on research combining multiple attempts with formative feedback. Recall that Yourstone, Kraye, and Albaum (2010) embedded feedback within their multiple attempts framework but did not describe the feedback design or analyze its influence on their results. We found only one previous study has tackled this issue. In an introductory physiology course, quizzes with formative feedback and untimed, unsupervised multiple attempts resulted in significant gains on subsequent exam performance (Marden, Ulman, Wilson, & Velan, 2013). The lack of literature suggests a need for further research to demonstrate the effectiveness of combining formative feedback and multiple attempts on assessments. By engaging students through feedforward and allowing students the opportunity to improve their performance on an assessment (and thus demonstrating mastery of the learning objectives), the line between formative and summative assessment is blurred. This study seeks to explore several questions: Do students who need to take advantage of a second attempt do so? If they do, does their performance on the assessment improve? Knowing the answers to these questions will provide critical insight into the pedagogical practice of feedforward with multiple attempts. Our study explores the following hypotheses:

(1) Students who do not earn an A on their initial attempt take advantage of the multiple attempts

(2) Students who take advantage of the multiple attempts outperform students who do not take advantage of the multiple attempts

(3) Students' second or third attempt on the assessment outperforms their first attempt

Materials and Methods

Participants

The institution where this work was performed is a medium-sized university with a "selective" rating according to the U.S. News and World Report [INSERT CITATION IN NON-BLIND COPY]. Grading and time on task data was obtained from the Learning Management System between October 2017 and February 2018 for 36 students enrolled in a 9-week introductory general chemistry course (lecture and lab) taught in the asynchronous online modality. Researchers examined 644 assessment results. The course content and learning outcomes are aligned with the traditional lecture course also taught at this university. The emphasis in the course is centered on problem solving and real world applications.

Students enrolled in the studied sections were primarily non-traditional students, with approximately 50% having a U.S. military affiliation and an average age of 34 years old. Students enrolled in the course for a variety of reasons. Certain degree programs required the chemistry lecture and lab course. Other degree programs require a 100-level physical sciences elective, some requiring the lab credit as well.

All data were aggregated with no individual identification of students, ensuring confidentiality. The data for this work was collected after the conclusion of the courses; no control group was utilized. This work was reviewed by the institutional Internal Review Board and deemed exempt.

Assessment Design

The assessments for this course were administered through the learning management system (LMS). Each assessment question was pulled from a pool, with each pool aligned with a module learning objective. Questions were closed, typically multiple choice, numerical answer, or multiple dropdowns. Questions were presented one at a time. The assessment was timed, with one hour for completion, though students could save and resume. Assessment questions were automatically graded by the learning management system. The feedback option in the LMS was leveraged to provide specific and actionable feedback once at the conclusion of the attempt on the assessment. Correct answers were not provided at any point in the assessment process.

The feedback used in this study was based on upon 3 principles of high-quality feedback: specific, actionable, timely. The use of LMS-embedded feedback ensures its timely nature. Specific and actionable feedback was designed according to Huba & Freed’s (1999) characteristics while achieving the supportive language with specific recommendations suggested by Bayerlein (2014). Examples of feedback are provided in Table 1. Though not listed here, the feedback also includes references to specific course resources. Within the LMS, instructors also provided big picture feedback after the completion of the multiple attempts.

Aligned with a well-regarded feedback philosophy, the feedback in this course is viewed as integral to the teaching and learning process and the assessment is viewed as a mechanism to enhance learning (Hattie & Timperley, 2007). Each assessment began with a short paragraph explaining when feedback will be provided and how to best use this feedback (feedforward) on the multiple attempts.

[INSERT TABLE 1 NEAR HERE]

The lecture quizzes constituted 40% of the overall course grade, with each quiz worth 4.44% of the overall course grade. The lab quizzes were worth 30% of the overall course grade, with each quiz worth 3.33% of the overall course grade. For lecture quizzes, students were permitted to complete the assessment twice, with no penalty for stopping after the first attempt. For pre-laboratory quizzes, students were permitted to complete the assessment three times. As a safety measure, students were required to pass the pre-lab quiz (>60%) in order to participate in the laboratory activities. For both lecture and lab, the highest score was awarded as the final grade on the assessment. The option for multiple attempts was communicated to students in many ways. The introduction to each assessment informed students of this option. The LMS also indicated the option for multiple attempts while engaged in the assessment. The course syllabus also informed students of this option. An announcement during the first week of the course reminded students and instructors also provided feedback to students throughout the term if they were not taking advantage of the multiple attempts but could potentially benefit from them because their initial attempt did not demonstrate full mastery of the content from that unit.

Statistical Analysis

Data testing was executed using StatCrunch Data Analysis on the Web and StatDisk (Triola, 2013). The first hypothesis was tested using Chi Square ($\alpha = .05$) at the appropriate degrees of freedom (Gay, Mills, & Airasian, 2006). The second hypothesis was tested using a one tailed t-test ($\alpha = .05$) for the lecture assessment data and ANOVA for the lab assessment data. The lab data was further explored using the post-hoc Tukey HSD tests. The third hypothesis was tested using paired samples t-test ($\alpha = .05$) and the fourth hypothesis was tested using a one tailed two sample t-test ($\alpha = .05$). The final three hypotheses were tested using regression analysis (Gay et al., 2006).

Results

Student Utilization of Multiple Attempts

The first hypothesis tested in this study explored if students who did not initially earn an A persisted by completing additional attempts. Initial attempt assessment scores that were scored 90% or above were not included in this analysis. Data for the remaining lecture and lab quiz scores were examined. All data were analysed using Chi Square test for equal expected frequencies for the quiz scores and pre-lab scores, tested separately (Table 2).

With $\alpha = 0.05$, we reject the null hypothesis for both lecture and pre-labs quizzes. There is evidence to support that students who do not earn an A defined as $\leq$ 90% are more likely to take the lecture assessment again. For the pre-lab assessment where three attempts were allowed, there is evidence to support that students who do not earn an A on their first attempt will try a second time. There is also evidence to support that students who do not earn an A after their second attempt will try a third time.

To be conservative because two statistical tests were performed on the same data set, we applied a Bonferroni correction ($\alpha = 0.025$) to the analysis of data for pre-lab quizzes where student scores were <90% after a second attempt. Under these conditions, a statistically significant difference was identified.

[INSERT TABLE 2 NEAR HERE]

We found no relationship between the week of the term and the number of students utilizing a second attempt on lecture or lab quizzes. The evenness of the participation in multiple attempts across the nine week term was also explored using the Chi Square test of good fit. With p = 0.6960 for the lecture data and, p = 0.9879 for lab data, the differences in utilization of two or three attempts was not significant across the term. Student use of multiple attempts was consistent throughout the term.

Impact of Multiple Attempts on Student Scores

The second hypothesis in this study sought to test if students who took advantage of the multiple attempts on the lecture and lab assessments outperformed students who did not take advantage of multiple attempts. The final scores of group were compared. This hypothesis was tested for the lecture quizzes using a one-tailed t-test ($\alpha = 0.05$). The resulting p-value was 0.7534 (Table 4). We fail to reject the null hypothesis, indicating that final scores were not significantly different between students who used one attempt compared to those who used multiple attempts. The similarity in final scores between the two groups is also evident in histograms (Figure 1). The means of the groups were similar, with students who did not retake the lecture quiz demonstrating a normal mean of 74.71% (standard deviation = 23.696) and students who retook the assessment having a normal mean of 72.322 (standard deviation = 19.369). While the average was slightly lower for those who retook the assessment, this difference is not statistically significant.

[INSERT FIGURE 1 NEAR HERE]

Because of the three allowed attempts, this hypothesis was tested for the pre-lab quizzes using 1-way Analysis of Variance (ANOVA), resulting in a p-value of 0.8630 (Table

3). As with the lecture quizzes, we fail to reject the null hypothesis. Final score averages for all three groups (those who took the quiz once, twice and three times) were between 86% and 87% (n=147). Students who retook the quizzes once or twice did not have a significantly higher score than those who took the quiz once. A post-hoc Tukey Honestly Significant Difference (HSD) tests was run to determine which groups of scores were significantly different from each other but showed no difference.

[INSERT TABLE 3 NEAR HERE]

The third hypothesis investigated whether a student's re-take outperformed their first attempt. To investigate this for the lecture quiz, a paired samples t-test was performed, resulting in a p-value of <0.0001 (Table 4). For the lecture quizzes, we reject the null hypothesis, with follow-on attempts tending to outperform first attempts on assessments by an average of 8.8 percentage points. First attempt scores averaged 59% and a median of 62.14. Second attempt average scores were 67.8% with a median of 70 (n = 119).

The chemistry lab scores were also evaluated to determine if scores improved from the first to second attempt and from the second to the third attempt. Students averaged 66.4% with a median of 66.3 on their first attempt. This score showed some improvement to 72.2% with a median of 71 on the second attempt. Third attempt average scores were significantly higher at 81.8% with a median of 83 (n = 55). The Analysis of Variance test results yielded a significant finding.

[INSERT TABLE 4 NEAR HERE]

Because of the significant findings from the ANOVA analysis, a post-hoc Tukey HSD test was run to determine which groups of scores were significantly different from each other for the students who attempted quizzes 3 times. All averages scores on attempts were statistically different, with the greatest difference seen between the first and the third attempt

(Tables 4 and 5). For the pre-lab quizzes, we reject the null hypothesis, with follow on attempts tending to outperform first attempts on assessments.

[INSERT TABLE 5 NEAR HERE]

Time on Task

Our study demonstrated that students who use multiple attempts use more time on both the lecture and lab assessments. Using a two-sample t-test to investigate the lecture quizzes, we found a p-value of <0.0001 (Table 6). Students who used two attempts on lecture quizzes (average 126 minutes with a standard deviation of 63.51 minutes) spent twice as much time as those who only used one attempt (average 62 minutes with a standard deviation of 30.23 minutes).

[INSERT TABLE 6 NEAR HERE]

The pre-lab quiz scores showed a similar pattern. The Analysis of Variance yielded a significant finding (0.0002) at α=.05 (Table 6). Students who took the quiz only once averaged 47.66 minutes on the task with a standard deviation of 37.10 minutes. Those who took the quiz twice spent 78.34 minutes on the task (standard deviation of 44.33 minutes) and those taking the quiz three times spent over 97 minutes on the assignment (standard deviation of 73.1 minutes). Significant differences for time on task existed between those who took the quiz once and those who took it two (p = .0308) or three (p < .0001) times. There was not a significant difference in time on task between students who took the quiz twice compared to students who took the quiz three times. There is evidence to support the idea that students who used two or more attempts spent more time on the assessment than those who used one attempt.

Levene’s test for homogeneity yielded a significant value of p = 0.0024, suggesting that the variances of the three groups of data compared for lab quizzes were significantly different. Due to that finding, a Kurskal-Wallis non-parametric test yielded a statistically significant result. ($p < 0.0001$) These additional tests were run to ensure results reported in the ANOVA and post-hoc Tukey did not lead to rejecting the null erroneously.

Having established that time on task in our study increased with multiple attempts, we probed whether time on task correlated to the grade earned on the 1st attempt (Figure 2), regardless of whether multiple attempts were completed. For the lecture quizzes, regression analysis showed a slight negative correlation, indicating a very weak relationship between time spent completing the first attempt on the assessment and the score on that assessment ($n = 182$). The Pearson's r correlation coefficient is -0.146. The Coefficient of determination (R^2) of 0.0214 indicates that this model explains less than 3% of the variation between the variables of time spent on the assignment and first attempt score. Over 97% of the variation can be due to other variables. The finding is not strong enough to state that time is a good predictor of lecture quiz scores on the first attempt.

Pre-lab scores ($n=147$) were similar to lecture quiz scores. The Pearson's r correlation coefficient was -0.0662 (Figure 2). With an R^2 of 0.004, the regression model explained less than 1% of the variation between time on task and 1st pre-lab quiz score and is thus not a good predictor.

[INSERT FIGURE 2 NEAR HERE]

With no correlation between time on task and first attempt score, we explored the correlation between time on task and the final grade earned on the assessment. For lecture quiz data ($n=182$), regression analysis demonstrated no discernible relationship, with a Pearson's r correlation coefficient of -0.107 and a coefficient of determination (R^2) of 0.012 (Figure 3). This model explains less than 2% of the variation between variables of time spent on task and the final lecture quiz score.

Similar results were found for the relationship between pre-lab final scores ($n=147$) and time on task. The Pearson's r correlation was -0.019, with an R^2 of 0.0004, meaning the regression model explained less than 1% of the variation between time on task and final score.

[INSERT FIGURE 3 NEAR HERE]

Discussion

Student Utilization of Multiple Attempts

Students who did not earn an A tried again

Students who did not earn an A on their initial attempt for a lecture or lab quiz took advantage of multiple attempts. Students completing the pre-lab quiz were permitted three attempts, yet this phenomenon held true whether the student failed to earn an A on the first attempt or the second attempt. This analysis does not explore the likelihood of a student to perform a second or third attempt to improve on an existing A. The existing literature on multiple attempts reports a range of utilization from 35% - 95% (Luebben, 2008; Orchard, 2016; Stewart, Panus, Hagemeier, Thigpen, & Brooks, 2014)

Student use of multiple attempts was consistent throughout the term

There was no relationship between the week of the term and student utilization of multiple attempts on either the lecture or the lab assessments. One could hypothesize that there would be a spike in the second week of the course as students become more aware of the option of multiple attempts on the assessments. We might also hypothesize a lag at the end of the term as students are more confident in their final course grade and can make informed cost-benefit analyses on the time investment of multiple attempts. In fact, other researchers have reported findings that support these hypotheses (Rhodes & Sarbaum, 2015; Stewart et al., 2014).

Impact of Multiple Attempts on Student Scores

Final scores are similar whether one attempt or multiple attempts were used

Students who use multiple attempts earn final assessment grades that are not statistically different from those who only use one attempt. The data shows only a minor difference of 1.69% on the final grade of lecture quizzes between students who performed multiple attempts and those who did not. The difference was even smaller between the groups of data for pre-lab quiz final scores which were about 86% to 87% with a difference of less than 1.3%.

Supporting our findings, Orchard (2016) reported that the difference in mean scores between single attempt quiz takers and multiple attempts used are negligible if not counterproductive. In their study, the outcomes were broken down between the module test results and by the range of score, showing the greatest difference between single and multiple attempts occurring in the very low student performance range between passing and not. While Orchard (2016) used a different statistical technique to explore this question for lecture data, both tests assume unequal variance and are appropriate tests of the hypothesis.

While evaluating the impact of multiple attempts in online economics class, the class that had 2 attempts on homework had significantly higher homework scores of 4-15% across 10 assignments (Rhodes and Sarbaum 2015, 120-121). However, this study also reported that multiple attempts on the exams in this course did not yield statistically significant differences in the final assessment score between the two groups. The multiple attempts concept was also evaluated at the middle and secondary school level (grades 5-12) during standardized end of course tests of mathematics and English, finding that the assessment scores were not significantly higher in students provided multiple attempts in either subject area (Stevens, 2013). Both of these studies are distinctly different from ours, though, as students in our study self-selected to complete only 1 attempt. All students had the option to complete multiple attempts. It is logical to assume that students who did not demonstrate mastery initially are more inclined to try again. It also follows that with repeated attempts the average student performance will be similar in both groups, those who used one attempt and those who used multiple attempts.

Students who use multiple attempts improve their score

Students' second attempt (or third attempt in the case of lab students) on an assessment tended to improve from their first attempt. This does not allow the conclusion to be drawn that a students who use second attempts perform better than those who only used one attempt (as explored in H_c). Instead, we conclude that a second attempt, when utilized, tends to yield a higher score than the first attempt.

The difference in the mean between the first and last attempt for all students using more than one attempt was significantly higher by an average of 8.8% for quizzes and 15.0% for lab scores. To further investigate, we explored the outcomes for students who did try again. For lecture quizzes, 71.5% of students who used a second attempt benefitted from doing so, with the average increase of 17.7 points (out of 100 points, standard deviation = 7.9 points). For the pre-lab quizzes, 87% of students who used multiple attempts benefitted from doing so. The average change in their final pre-lab score was 23.5 points (out of 100 points, standard deviation = 6.1 points). While this is a remarkable improvement, this value could be skewed by the concept of the "throw-away attempt" where a student's first attempt is exploratory and not a good faith attempt to demonstrate mastery.

On average, the number of students earning an A on the lecture quizzes increased by 12% and those who transitioned from a failing grade to a passing grade increased by 19%. Interestingly, these averages were duplicated in the lab course, with a 12% increase in As and a 19% increase in passing final scores.

Our data aligns with the findings of Rhodes and Sarbaum (2015), who reported that homework scores increased on multiple attempts. However, they noted that the scores did not tend to increase when using multiple attempts on summative exams; one exam showed a decrease in scores on the second attempt while the second exam showed a slight increase in scores on the second attempt.

An exploration of ungraded self-testing for pharmacy doctoral students found that graded exam scores were higher following implementation of practice quizzes for 3 of 4 testing periods (Stewart et al., 2014). They also found that neither PCAT nor GPA correlated with high exam scores, lending support to the argument that the self-testing benefitted student performance by removing "underlying intelligence" as an influencing variable. They emphasize the importance of the self-test in allowing students to gauge their mastery objectively, improving metacognition and allowing the student to seek intervention or adjust study techniques. While this study has marked differences to our work presented here, it provides further peripheral support to our findings.

Time on Task

Students who use multiple attempts spend more time on the assessment

We found that students who use multiple attempts invest more time on the assessment. This finding is supported by Rhodes and Sarbaum (2015), who reported that students given the option of a second attempt on homework yielded qualitatively similar or higher time spent on homework. Furthermore, they found that time on each attempt tended to increase, with the first attempt as much as 12-15 minutes less that future attempts.

Time on task does not correlate to score

The relationship between number of attempts and time on task is fairly obvious; of more interest is whether an increased time investment is correlated to a higher final score. Regression analysis demonstrated that this model does not explain the variation between a student's score on their first attempt on the lecture or lab assessments and the time they spent on the task. Furthermore, no relationship was revealed between the total time on task and the final grade for both lecture and lab assessments. Interestingly, this conflicts with the time-on-task hypothesis proposed in 1963 (Carroll, 1963). However, the measure of time on task in this study was limited to the duration of the online assessment and did not capture preparation time prior to completing the assessment. Recent literature, while supporting the positive correlation between time on task and learning outcomes, did find large variability (Godwin et al., 2016).

The exploration of the relationship between time on task and final score could be influenced by the potential for “throw-away attempts” where students use an attempt on the assessment to simply preview the assessment before completing a good faith attempt. Previous researchers have claimed that the similarity between time spent overall on assessments with or without multiple attempts in combination with reduced time spent at first, when multiple attempts were available are indicative of the "throw away attempt" (Rhodes & Sarbaum, 2015). However, we feel that identifying a “throw-away attempt” requires making significant assumptions and is not justified without qualitative data to support the conclusion (e.g. student survey or interviews).

Limitations of This Study

There are several limitations to address in this study. The primary limitation of this study was the approach using existing data. Because this study was a retrospective investigation of the effects of multiple attempts in a course, there is no true control. There may be ethical concerns with establishing a control group, given the evidence in this secondary data.

Additionally, this study was unable to fully control all moderating variables. Students may or may not have used the feedback prior to initiating an additional attempt on the assessment. Furthermore, the opportunity to use multiple attempts does not require students to perform to the utmost of their ability until the final attempt and therefore early attempts may not reflect an accurate depiction of their best work on each attempt. Gains in student mastery may be artificially inflated if initial attempts are not good faith efforts.

The amount of time spent on the assessment is highly variable. There is no way to determine how much of the time on task recorded automatically through the learning management system was actually dedicated to the assessment. It is possible that the student started the assessment and then navigated to other tasks for unknown amounts of time. Furthermore, any time spent preparing for the assessment off-line was not measured.

This study only explored one subject area, which limits the generalizability to other populations. While both the lecture and lab for an introductory physical sciences course showed similar results for all hypotheses tested, it would be informative to explore these research questions in additional physical science disciplines and beyond.

Conclusions

The data from this study indicated that students who did not earn an A tried again and that those who did try again tended to do better at demonstrating mastery of the subject (evidenced by a higher assessment score). However, the average scores were similar between students who only used on attempt and those who used multiple attempts. Because of “throw-away attempts” and the disincentive for a good faith effort on the initial attempt, our results regarding multiple attempts are not a wholly accurate depiction of effective "learning gains" but rather, the demonstration of an effective pedagogical technique to provide feedback and an application opportunity, thus scaffolding the assessment for student success.

While multiple attempts required an increased time investment from students, it did not correlate to a higher final score on the assessment. Based on the data presented, it is fair to say that multiple attempts only give the opportunity for a student to be statistically better at demonstrating mastery on the assessment. While this pedagogical choice requires an up-front time investment in course design in building robust question pools for assessments that include high quality embedded feedback, this pedagogical choice may close the gap between instructor expectations and student performance, with the noted benefit for initially-underperforming students.

Future studies should explore if and how students utilized the feedback through a qualitative survey. A survey could also explore the idea of a "throw away attempt". The current study should be replicated in additional STEM courses to ensure generalizability of results.

References

Baker, D. J., & Zuvela, D. (2013). Feedforward strategies in the first-year experience of online and distributed learning environments. *Assessment & Evaluation in Higher Education, 38*(6), 687-697. doi:10.1080/02602938.2012.691153

Bayerlein, L. (2014). Students' feedback preferences: How do students react to timely and automatically generated assessment feedback? *Assessment & Evaluation in Higher Education, 39*(8), 916-931. doi:10.1080/02602938.2013.870531

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64,* 723-733.

Cheng, Y., & Cheng, L. (2016). A short note on the relationship between pass rate and multiple attempts. *Journal of Educational Measurement, 53*(4), 431-447. doi:10.1111/jedm.12124

Coates, H., James, R., & Baldwin, G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary Education and Management, 11*(1), 19-36. doi:10.1007/s11233-004-3567-9

Dulama, M. E., & Ilovan, O. (2016). How powerful is feedforward in university education? A case study in romanian geography education on increasing learning efficiency. *Kuram Ve Uygulamada Egitim Bilimleri, 16*(3), 827-848. doi:10.12738/estp.2016.3.0392

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: Are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice, 24*(1), 34-39. doi:10.1111/emip.12059

Gay, L. R., Mills, G. E., & Airasian, P. W. (2006). *Educational research: Competencies for analysis and applications* (8th ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Gaytan, J., & McEwen, B. C. (2007). Effective online instructional and assessment strategies. *American Journal of Distance Education, 21*(3), 117-132. doi:10.1080/08923640701341653

Godwin, K. E., Seltman, H., Almeda, M. V. Q., Kai, S., Baker, R. S., & Fisher, A. V. (2016). The variable relationship between on-task behavior and learning. *Learning and Instruction, 44,* 128-143. Retrieved from http://www.upenn.edu/learninganalytics/ryanbaker/Godwin_Cogsci_2016_Final.pdf

Goldsmith, M. (2008, Try feedforward instead of feedback. *The Linkage Leader, ,* 1-5. Retrieved from http://www.linkageanz.com.au/uploads/pdf/Marshall_Goldsmith_Try_Feedforward_Instead_of_Feedback_1102%5B1%5D.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Huba, M. E., & Freed, J. E. (1999). *Learner-centered assessments on college campuses: Shifting the focus from teaching to learning.* Needham Heights, MA: Allyn & Beacon.

Hughes, G. (2011). Towards a personal best: A case for introducing ipsative assessment in higher education. *Studies in Higher Education, 36*(3), 353-367. doi:10.1080/03075079.2010.486859

Koen, K., Bitzer, E. M., & Beets, P. A. D. (2012). Feedback or feedforward? A case study in one higher education classroom. *Journal of Social Sciences, 32*(3), 231-242. doi:10.1080/09718923.2012.11893068

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Journal of Psychological Science, 23*(22), 1337-1344. doi:10.1177/0956797612443370

Luebben, A. J. (2008, 11/1). Offering multiple test trials: Educational folly or learning opportunity? *Online Cl@ssroom,* Retrieved from http://augmenting.me/cte/resources/newsletters_archive/OC0811.pdf

Marden, N. Y., Ulman, L. G., Wilson, F. S., & Velan, G. M. (2013). Online feedback assessments in physiology: Effects on students' learning experiences and outcomes. *Advances in Physiology Education, 37*(2), 192-200. doi:10.1152/advan.00092.2012

Orchard, R. K. (2016). Multiple attempts for online assessments in an operations management course: An exploration. *Journal of Education for Business, 91*(8), 427-433. doi:10.1080/08832323.2016.1256262

Rhodes, M. T., & Sarbaum, J. K. (2015). Online homework management systems: Should we allow multiple attempts? *American Economist, 60*(2), 120-131. doi:10.1177/056943451506000203

Rodriguez-Gomez, G., & Ibarra-Saiz, M. S. (2015). Assessment of learning and empowerment: Towards sustainable learning in higher education. In M. Peris-Ortiz, & J. Merigo Lindahl (Eds.), *Sustainable learning in higher education* (pp. 1-20). Cham: Springer. doi:10.1007/978-3-319-10804-9

Sewell, J., Frith, K. H., & Colvin, M. M. (2010). Online assessment strategies: A primer. *Journal of Online Learning and Teaching, 6*(1), 297-305.

Stevens, S. L. (2013). *The impact of the multiple attempts at mastery philosophy on the academic achievemnt and behavior of elementary school students* (Ed.D.). Retrieved from http://search.proquest.com.ezproxy.libproxy.db.erau.edu/docview/1586074901?accountid=27203

Stewart, D., Panus, P., Hagemeier, N., Thigpen, J., & Brooks, L. (2014). Pharmacy student self-testing as a predictor of examination performance. *American Journal of Pharmaceutical Education, 78*(2), 165. doi:10.5688/ajpe78232

Stodberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education, 37*(5), 591-604. doi:10.1080/02602938.2011.557496

Triola, M. (2013). Statdisk [computer software] Pearson Education Inc.

Wiggins, G. (2012). Seven keys to effective feedback. In M. Scherer (Ed.), *On formative assessment: Readings from educational leadership* (pp. 24-35). Alexandria, VA: ASCD.

Retrieved from https://books.google.com/books?id=ycWqDAAAQBAJ&dq=timely+feedback&lr=

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.

Wolkowitz, A. A. (2011). Multiple attempts on a nursing admissions examination: Effects of the total score. *Journal of Nursing Education, 50*(9), 493-501. doi:10.3928/01484834-20110517-07

Yourstone, S. A., Kraye, H. S., & Albaum, G. (2010). Online quantitative-based assignments - are more attempts better for learning? *Decision Sciences Journal of Innovative Education, 8*(2), 347-351. doi:10.1111/j.1540-4609.2010.00260.x

Table 1: Select feedback for the first quiz in an introductory chemistry course

Topic	Question	Feedback
Significant Figures	Which number below contain 3 significant figures?	Zeroes to the left of the nonzero digits are never significant. Zeroes in between nonzero numbers are always significant. Zeroes to the right of nonzero numbers are significant if there is a decimal present.
Dimensional Analysis	Determine the number of atoms across the diameter of a human hair given that the diameter of an atom is 0.1 nm and the diameter of a human hair is 0.1 mm.	This can be solved two ways. If you're comfortable with the prefixes and scientific notation, you can just move the decimal accordingly. You can also write out the conversion factor to get between nm and mm. Be sure to write out your dimensional analysis so that you can ensure your units cancel out.
Subatomic Particles and Atomic Models	How should this diagram be changed to properly represent Lithium - 8?	Which subatomic particles change to form isotopes?
Properties of Matter	Which of the following represents a chemical property of copper metal?	The observation of a chemical property changes the identity of the substance.
Classifying Matter	Which of the following is a homogeneous mixture?	Homogeneous mixtures have uniform appearance and composition.

Table 2: Chi square Goodness of Fit table for Hypothesis 1 (α = 0.05)

Assessment	Variable	n	Retook	<90% Did not retake	DF	Value	p-value
Lecture Quiz	Grade < A	160	119	41	1	38.0250	<0.0001
Pre-Lab Quiz	Grade <A after initial quiz	126	108	18	1	64.2857	<0.0000
	Grade still <A after 2nd attempt	82	58	24	1	14.0976	.0002*

Note. *Bonferroni corrected alpha level of .025 (Triola, 2018).

Table 3: Statistical analysis of Hypothesis 2 (α = 0.05)

Assessment	n	Statistical Test	p-value
Lecture Quiz	182	One-tailed t-test	0.7534
Lab Quiz	147	ANOVA	0.8630

Table 4: Statistical analysis of Hypothesis 3 (α = 0.05)

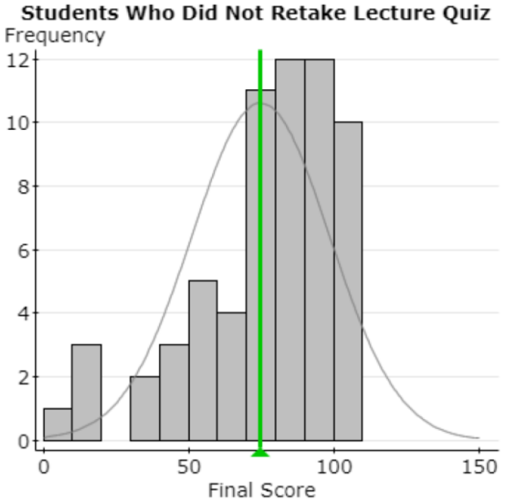
Assessment	n	Statistical Test	p-value
Lecture Quiz	119	Paired samples t-test	<0.0001
Lab Quiz	58	ANOVA	<0.0001
Attempt 1 vs Attempt 2		Tukey HSD	0.0599
Attempt 1 vs Attempt 3		Tukey HSD	<0.0001
Attempt 2 vs Attempt 3		Tukey HSD	0.0005

Table 5. Mean comparison of Lab Scores for Three Attempts (n=58)

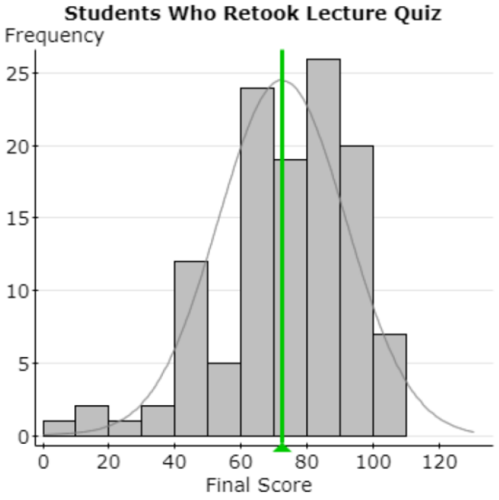
Assessment	Mean Pre-Lab Quiz Score (%)
First Attempt	66.44
Second Attempt	72.19
Third Attempt	81.81

Table 6: Statistical analysis of Time on Task ($\alpha = 0.05$)

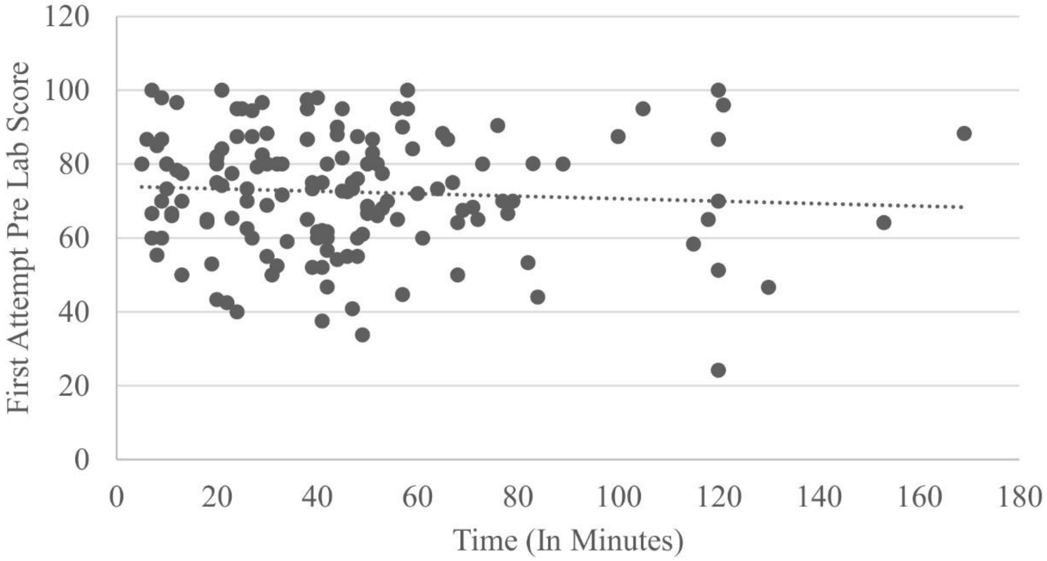
Assessment	n	Statistical Test	p-value
Lecture Quiz	182	Two-samples t-test	<0.0001
Lab Quiz	147	ANOVA	0.0002
		Levene’s test	0.0024*
		Kurskal-Wallis	<0.0001*
1 Attempt vs 2 Attempts		Tukey HSD	0.0308
1 Attempt vs 3 Attempts		Tukey HSD	<0.0001
2 Attempts vs. 3 Attempts		Tukey HSD	0.1851

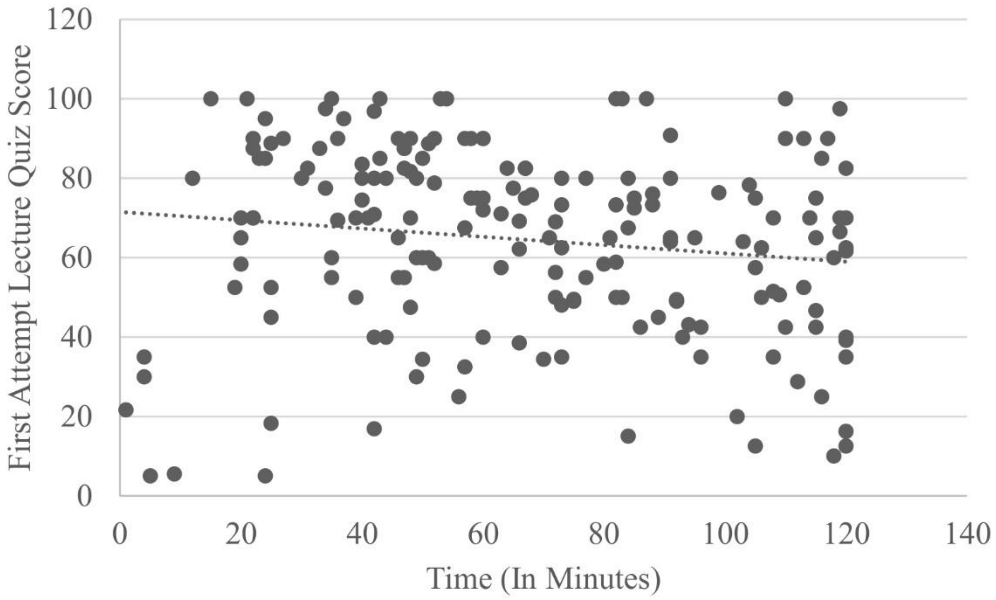


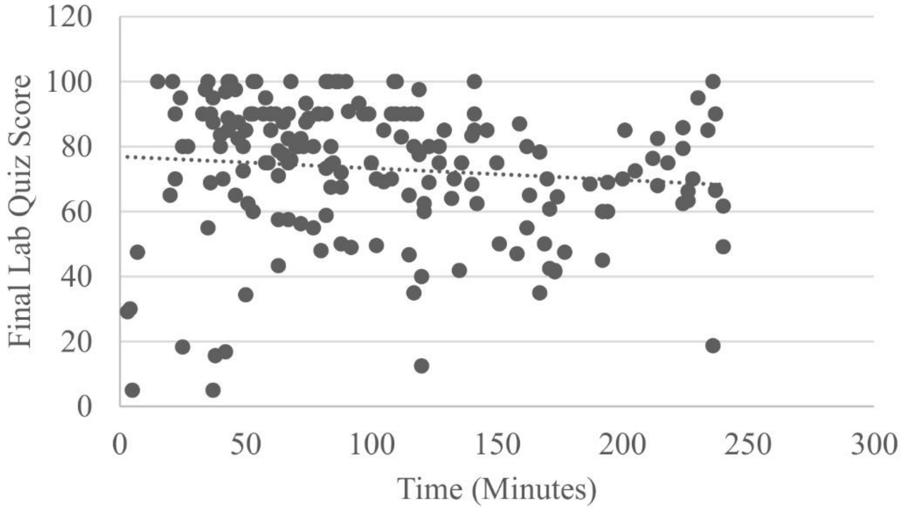
Students Who Did Not Retake Lecture Quiz



Students Who Retook Lecture Quiz


First Attempt Pre Lab Score
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
Time (In Minutes)


120
100
80
60
40
20
0
First Attempt Lecture Quiz Score
0
20
40
60
80
100
120
140
Time (In Minutes)


Final Lab Quiz Score
120
100
80
60
40
20
0
0
50
100
150
200
250
300
Time (Minutes)

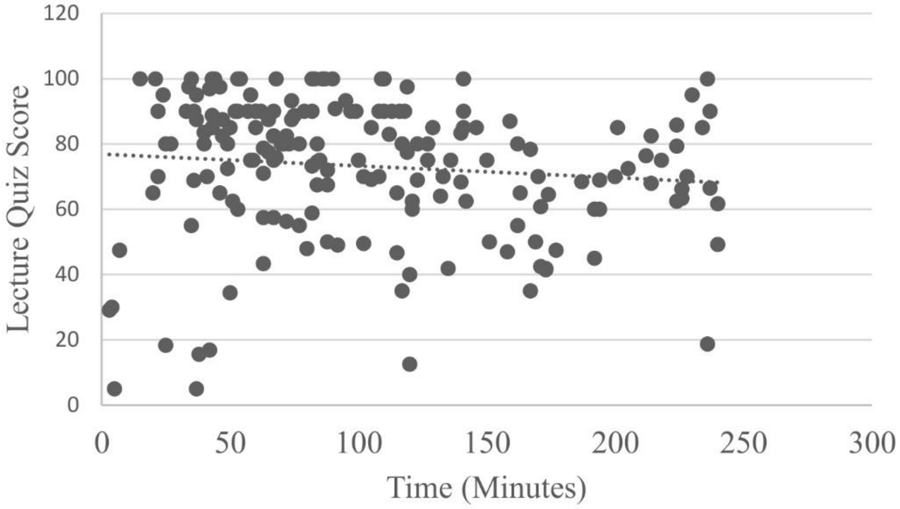

Lecture Quiz Score
120
100
80
60
40
20
0
0
50
100
150
200
250
300
Time (Minutes)

Figure 1: Histograms to compare final scores based on single vs. multiple attempts

Figure 2: Relationship between the a) pre-lab and b) lecture quiz first attempts and time on task

Figure 3: Relationship between time on task and final grade for a) pre-lab and b) lecture quizzes