ANOVA Analysis of Student Daily Test Scores in Multi-Day Test Periods

Matthew L. Mouritsen, PH.D.
Department of Accounting and Taxation
Weber State University
Ogden, Utah

Jefferson T. Davis, PH.D., CPA, CISA
Department of Accounting and Taxation
Weber State University
Ogden, Utah

Steven C. Jones
Institutional Research
Weber State University
Ogden, Utah

ABSTRACT

Instructors are often concerned when giving multiple-day tests because students taking the test later in the exam period may have an advantage over students taking the test early in the exam period due to information leakage. However, exam scores seemed to decline as students took the same test later in a multi-day exam period (Mouritsen and Davis, 2012). This study reports mean test score analysis of a four-day exam period. Students with higher cumulative GPAs tend to take the exam earlier in the testing period. The majority of students take the exam the last day of the testing period. Test score variance for each test day also increases with each test day. One-way ANOVA analysis finds that mean test scores of students who take the test later in the test period significantly decline. Pairwise comparisons that assume unequal numbers of observations in each group as well as unequal variances of exam scores for each day, show that day 4 mean scores are significantly less than days 1, 2, and 3. The only other pairwise difference is day 1 and day 3. Further, a 4 X 2 (4 test days by two different professors) ANCOVA analysis is also reported where cumulative GPA and Test # (4 or more tests each semester) are used as control variables to see if student test scores still decrease for students taking the test later in the testing period. The results show significant decreases in mean test scores as students take the test later in the testing period even when controlling for students' cumulative GPA and Test # within the semester. An estimated marginal means analysis further shows that the upper bound of day 4 is below the lower bound of days 1, 2, and 3, consistent with pairwise comparisons of test score means. The results suggest that information leakage, if any, is not evident in multi-day test scores. The results suggest that an instructor may have an opportunity to further help students taking the exam later in the exam period. Further research on demographics, test preparation, procrastination, self-efficacy, and emotional intelligence of students taking multi-day tests is in order (Hen and Goroshit, 2014).

INTRODUCTION

Many universities are using testing centers to allow students to take tests when it is more convenient for the student. One of the issues related to testing centers in general, and specifically for tests that can be taken by students over multiple days, is the risk of information leakage to students who take that test later in the test period. However, two studies have found that instead of test scores being higher for students taking the test late in the multi-day testing period, test scores are actually lower for students who take the test later in the multi-day testing period (see Mouritsen and Davis, 2012, and Reed and Holley 1989). Although, this information does not mean information leakage does not take place, it does suggest that other factors are much more prominent in determining test scores in a multi-day test period than any information leakage that may take place. For example, there are several articles in the education literature that study procrastination in academic settings.

The objective of this research is to discover why average test scores of students who take the test at the end of a multi-day testing period are lower than average scores of students who take the test earlier in the testing period.

This study analyses test scores of students taking exams over multi-day testing periods for introductory financial accounting (Accounting 2010) and introductory managerial accounting (Accounting 2020) courses taught by two different instructors over several semesters. The tests were all administered in the testing center over a 4-day period. Students were allowed to select when to take the test during the 4-day testing period. The exams were all multiple choice and no time limit was given. The analyses in this study include test scores from different tests taken during different semesters. Exhibit 1: shows the Distribution of Students included in the study taking the tests during each of the successive four test days. The data includes only tests where four test days were used so that the test percentages for each course could be consistent based on the number of days. Exhibit 1: Distribution of Students Taking Exam Each Day for Both Courses shows that more students took the test each successive day of the test period and the total number of tests included in the study for each course. The total number of tests did include up to four test scores from each individual student for different exams taken during a semester. Exhibit 2: Distribution of Mean Exam Scores by Test Day for Both Courses shows that test percentage scores drop with each successive day of the test period. One might expect that better students tend to take the test earlier in the exam period. Exhibit 3: Mean GPA of Students by Test Day for Both Courses shows that, in fact, the average cumulative GPA of students who take the test earlier is higher than the average GPA of students who take the test later. This research is thus aimed at discovering and analyzing what other course and student characteristics might play a role in students' test taking and scores over a multi-day testing period.

Student characteristics were also paired with the test scores of each student as well as information about what day the test was taken by each student during a 4-day testing period. In addition to student test percentage scores, the student test percentages were matched with other test information and student characteristics, including exam number during the semester the course was taken, class level (freshman, sophomore, etc.), whether the student was full-time or part-time, and age of student.

RESEARCH DESIGN AND HYPOTHESES

The descriptive statistics support the finding that students' average test scores get worse by day as the multi-day testing period progresses. With the ultimate objective of this research being to discover why average test scores of students who take the test at the end of a multi-day testing period are lower than average scores of students who take the test earlier in the testing period, this research takes the following basic approach: First an ANOVA model is used to determine whether there are differences overall in the mean test scores for each of the four days in the testing period. Then, if an overall difference is found, a pairwise test is used to determine which test days exhibit different mean test scores from each of the other test days. Statistical correlations are also run to find relationships between mean student test scores and various course and student characteristics. Using the information from these correlations an ANCOVA model is developed to test whether these course and student characteristics are statically significant variables for determining mean test score by test day. Finally, a marginal means analysis is used to further study the relationship of these student characteristics to the day they took the test and the mean test score for each day.

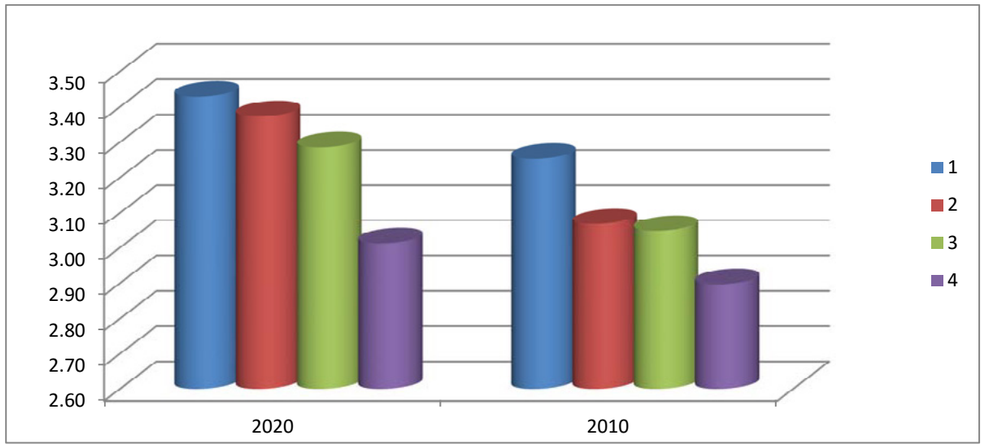
Exhibit 1
Distribution of Students Taking Exam Each Day for Both Courses

Course	Item	Day 1	Day 2	Day 3	Day 4	Total
Accounting 2010	**# of Students**	106	147	154	310	717
	% of Students	15%	21%	21%	43%	100%
Accounting 2020	**# of Students**	68	60	184	428	740
	% of Students	9%	8%	25%	58%	100%

Exhibit 2
Distribution of Mean Exam Scores by Test Day for Both Courses

Course	Item	Day 1	Day 2	Day 3	Day 4	Overall Average
Accounting 2010	Mean score by day	88%	85%	82%	72%	79%
Accounting 2020		80%	76%	79%	71%	74%
Combined		85%	82%	80%	71%	77%

Exhibit 3
Mean GPA of Students by Test Day for Both Courses



Average of GPA					
Course	Day 1	Day 2	Day 3	Day 4	Overall Average
2020	3.43	3.37	3.29	3.01	3.15
2010	3.25	3.07	3.05	2.90	3.02
Average by Day	3.32	3.16	3.18	2.96	3.08

ANOVA Hypothesis and Test Results

To determine whether the mean test score (test percentage) differs overall for the 4-day test period an ANOVA model is appropriate. The ANOVA model provides an indication if the mean test scores for the four days are statistically different based on days. Formally, the null hypothesis is as follows:

ANOVA H1(null): No overall mean test score differences between test days exist.

If H1(null) is not rejected, then the results of the research end with the finding that, on average, it does not matter which day a student takes the exam in relation to their mean test score. If the null hypothesis is rejected, then the

results indicate that the mean test scores do differ by day of the test period. Based on the descriptive statistics found in Exhibit 3, the expectation is that the null hypothesis will be rejected, in other words, statistical differences exist in mean test scores for students taking the tests over a 4-day test period. One student characteristic that may seem somewhat obvious is that better students will take the test earlier in the test period. Exhibit 3 shows student GPA in relation to mean test score by test day. There may be other explanations for the results as well. Further analysis is in order if statistical differences are found using the ANOVA test.

The ANOVA to determine if statistical differences between mean test scores for the 4-day test period rejects the null hypothesis that there are no differences based on which day the test was taken by students. Exhibit 4 shows the descriptive statistics, the ANOVA and Brown-Forsythe results the test scores for the 4-day test period.

The mean (average) test scores in the descriptive panel match the means listed in Exhibit 3. The descriptive panel also provides the number of students taking the test in each of the four days, the standard deviation for each of the 4-days test scores, and the 95% confidence intervals for each of the 4-days test scores. The main result of the ANOVA procedure shows strong differences between the mean test scores for the four test days (significance of .000). An important aspect of the descriptive statistics reveals that many more students take the exam on the second day than on the first day. Day three and four have more students who take the exam than the previous days as well. Also notice that, with the exception of day three, the standard deviation (a measure of variation from the mean test score for the day) increases during the 4-day test period. It is not surprising that the standard deviation of test scores increases with the number of students taking the exam on a given day—more students, more variety. This finding suggests that students taking the exam each day may have differences that lead to different exam scores for each day. The fact that the number of students taking the exam each day increases by day and that standard deviations for each day test scores also generally increase suggests that the ANOVA may not be valid. ANOVA procedures generally assume homogeneous (similar) variances in the data. To test for non-homogenous (non-similar) variances, the Brown-Forsyth test was also performed. The Brown-Forsyth test results show statistical differences in mean test scores for the multi-day testing period even when accounting for unequal variances and unequal number of students taking the test each day. With statistical differences in mean test scores for the 4-day testing period confirmed by the ANOVA and Brown-Forsyth tests, the next step is to test for pairwise differences of mean test scores for each day.

Exhibit 4
Mean Test Percent Score
ANOVA Results and Brown Forsyth for Non-homogeneity of Variance

Descriptives

Day	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	174	.851935	.1332764	.0101037	.831993	.871877
2	207	.822888	.1571310	.0109214	.801356	.844420
3	338	.799901	.1418721	.0077168	.784722	.815080
4	738	.713808	.1708646	.0062896	.701460	.726155
Total	1457	.765773	.1674227	.0043862	.757169	.774377

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.354	3	1.451	57.835	.000
Within Groups	36.459	1453	.025		
Total	40.812	1456			

Robust Test of Equality of Means[1]

	Statistic[2]	df1	df2	Sig.
Brown-Forsythe	65.282	3	970.419	.000

[1]The Brown-Forsythe test, which accounts for the lack of variance homogeneity, indicates statistically significant results even with unequal variances and unequal number of test scores in each day.

[2] Asymptotically F distributed.

Pairwise Hypothesis and Test Results

In the case of differences, a pairwise comparison can provide information as to any statistical differences between mean test scores for each day in relation to each of the other days. Formally, the null hypothesis states:

Pairwise H2 (null): No day-to-day pairwise differences in mean test scores for each of the four test days exist.

If H2 (null) is rejected, we will then have information concerning which test days' mean test scores are statistically different from each of the other test days' mean test score.

Exhibit 5
Pairwise Comparisons of Test Days' Mean Exam Scores

Multiple Comparisons
Tamhane's T2 Pairwise Test1

Exam Day (a)		Exam Day (b)		Mean Difference (a-b)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
dimension2	1	dimension3	2	.0290471	.0148782	.272	-.010304	.068398
			3	.0520339*	.0127135	.000	.018404	.085664
			4	.1381275*	.0119014	.000	.106621	.169634
	2	dimension3	1	-.0290471	.0148782	.272	-.068398	.010304
			3	.0229868	.0133726	.418	-.012371	.058345
			4	.1090804*	.0126030	.000	.075735	.142426
	3	dimension3	1	-.0520339*	.0127135	.000	-.085664	-.018404
			2	-.0229868	.0133726	.418	-.058345	.012371
			4	.0860936*	.0099553	.000	.059834	.112353
	4	dimension3	1	-.1381275*	.0119014	.000	-.169634	-.106621
			2	-.1090804*	.0126030	.000	-.142426	-.075735
			3	-.0860936*	.0099553	.000	-.112353	-.059834

*. The mean difference is significant at the 0.05 level.
[1] The Tamhane's T2 is a pair-wise procedure based on the Student t-distribution. Tamhane's is a more conservative post hoc comparison for data with unequal variances and is appropriate when variances are unequal and/or when the sample sizes are different." (source: chapter 11, page 256 of Basic Statistics and Pharmaceutical Statistical Applications By James E. De Muth

The results of the pairwise test comparing the mean test score of each day to each of the other three days is found in Exhibit 5: Pairwise Comparisons of Test Days' Mean Exam Scores.

Pairwise procedures result in mixed results as to whether the null hypothesis of no means test score differences of a particular day in relation to each of the other days is rejected or accepted. The results show that day 1 mean score is not statistically higher than day 2 (272), but it is higher than the mean test scores of day 3 (.000) and day 4 (.000). The day 2 mean test score is not different than day 1 (.272) or 3 (.418), but it is higher than day 4 (.000). Finally, day 3 mean test score is higher than day 4 test score (.000). It should be noted that day 4 mean test score is significantly lower than each of the other three days' mean test scores (.000).

The Tamhane's T2 pairwise procedure was chosen because this particular pairwise test is appropriate when unequal samples sizes exist and when variances (i.e standard deviations) are also unequal. Since pairwise differences between mean test scores for most of the days are found, further analysis is needed to determine why the test scores for different test days tend to get lower as test days progress from day 1 through 4. Particularly, further analysis seeks to find answers to the question, "Why are test scores for the last day, day 4, lower than each of the other three days of the exam period?"

Exhibit 6
Correlations between Mean Exam Score as a Percentage and Other Variables (N = 1457)

Variable	Pearson Correlation	Significance (2-Tailed)
Cumulative GPA	.437 **	.000
Exam Day	-.312 **	.000
Exam #	-.292 **	.000
Class Level	.046	.079
Semester	.008	.756
Age	.006	.830

** Significant at .01 level (2-tail)

Class level exhibited some correlation with test scores (.045) but the significance level (.079) did not approach reach .01. Semester and student age had extremely weak correlations and were very far from statistical significance. Whether a student was full or part time also did not show a relationship with test scores. These correlations were then used to determine what variables would be used in the ANCOVA.

ANCOVA Hypotheses and ANCOVA and Marginal Means Tests Results

Correlations of Test Scores with Student and Course Characteristics

Since some pairwise differences between each days' mean test scores were found, the next step is to study potential reasons why different days in the testing period yield different mean test scores. Statistical correlation procedures are used to find strong or weak relationships between student and course characteristics (i.e. course/prof, test number, student GPA, class level, full/part time) and test scores. Exhibit 6 shows the Correlation results between student test scores and student's cumulative GPA, exam day, exam number, class level (freshman, sophomore, etc.), semester, and age of student.

The Pearson correlations were significant for GPA (.437; .000), exam day (-.312; .000), and exam number (-.292; .000). Exam number refers to the first to last exams in the semester. The correlation shows that exam scores tend to be lower for exams given later in the semester. This result makes sense as exams taken later in the semester typically deal with more difficult topics or topics that build on information from the earlier part of the course. And of course it makes sense that exam day has a negative correlation with text scores.

Based on the correlation results, an ANCOVA model was developed to see if mean test scores by test day still differ if these course and student characteristics are used as control variables in the ANCOVA model. In general, ANCOVA is a combination of ANOVA and linear regression. The ANOVA includes a dependent variable (mean test scores) with one or more categorical independent variables (4 test days and 2 different courses), combined with other control variables to "correct" for or take into account other variables or characteristics that may confound or make a difference in the predictive model. The ANCOVA model tests for statistical differences in mean test day scores while controlling for these characteristics. The ANCOVA results will also find which of these variables statistically contribute or help to explain differences in mean test scores for each of the four test days. The null hypotheses related to the ANCOVA are as follows:

ANCOVA H3A (null) No mean test score differences from main effects in 4X2 (4 days X 2 courses/professors) design.

ANCOVA H3B (null) Covariates (Student GPA, Test #) are not significant variables and do not contribute to any mean test score differences in relation to 4-day exam period nor 2 different courses/professors.

Finally, a marginal means test was conducted to explore further differences in any ANCOVA results to show the percentage of students within each day's mean scores.

The ANCOVA was a 4X2 design (4 test days by 2 courses/professors) for the main effects. The covariates included in the model to control for characteristics that might confound main effects on the ANOVA were student GPA and exam number.

The results show the H3A (null) and H3B (null) are both rejected. In other words, the main effects, test day and course/professor were significant contributing variables to predicting the test score. Also, the two covariates (student GPA, and exam #) were significant to the ANCOVA model. Therefore, even when controlling for student GPA and exam #, the main effect variables of test day and course/professor were still strong predictors of student test scores. The results also show that student GPA and exam # have an impact on student test scores. The interaction between test day and course did not, however, significantly impact the strength of the model in explaining student test scores. The ANCOVA results achieved an adjusted R^2 of .366. This means that, overall, the model explains student test scores fairly well.

The estimated marginal means further shows that day four test scores have an upper 95% confidence interval (upper bound is .743) that is lower than all the other days (lowest lower bound day 3 is .775) 95% confidence lower bound even when controlling for student GPA and Test number.

A 95% confidence interval means that with 95% probability, the true mean test score is within that interval. Since the upper bound of day four exam scores is lower than the lower bound of any other day's mean test score, it is clear that there is very small probability (5%) that day four mean test score overlaps any other days' true mean test score. The marginal means statistics resulting from the ANCOVA model show that the day four group characteristics in relation to exam scores are strongly different than students taking the test on the other three days. The day four group is the largest group, has the lowest average GPA, and the largest test score variation. Although the marginal means standard error is smallest for day four, the standard deviation for day four test scores is the largest (see Exhibit 4). The reason the marginal means standard error is smallest is largely due to the fact that the number of students who take the test on day 4 is much larger than the other three days. A higher number N typically strengthens the statistical ability to narrow the confidence interval.

Exhibit 7
4x2 ANCOVA Design
(4 levels:{day 1, day 2, day 3, day 4} x 2 levels: {professor 1, professor 2})
Covariates: Cumulative GPA, Exam #

Tests of Between-Subjects Effects
Dependent Variable: Exam Score Percent

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Cumulative GPA	6.786	1	6.786	381.698	.000
Exam#	3.419	1	3.419	192.302	.000
ExamDay	1.414	3	.471	26.515	.000
Professor	1.124	1	1.124	63.221	.000
ExamDay * Professor	.054	3	.018	1.009	.388
Error	25.725	1447	.018		
Total	895.209	1457			

R Squared = .370 (Adjusted R Squared = .366
All main effect and covariates are statistically Significant.
*No Statistically significant interaction effect between ExamDay and Professor

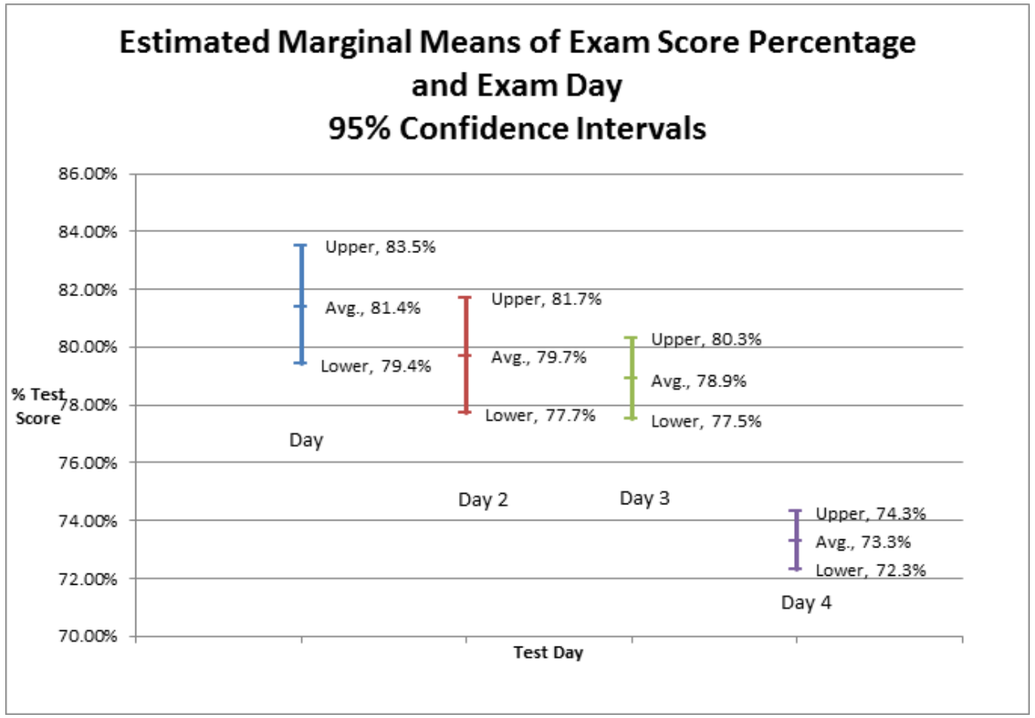
Exhibit 8

Estimated Marginal Means of Exam Score Percentage and Exam Day

Dependent Variable: Percent of Exam Score

Exam Day		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
dimension1	1	.814a	.010	.794	.835
	2	.797a	.010	.777	.817
	3	.789a	.007	.775	.803
	4	.733a	.005	.723	.743

a. Covariates appearing in the model are evaluated at the following values: CUM_GPA_UGRAD = 3083.18, Exam # = 2.62.



LIMITATIONS, SUMMARY, CONCLUSIONS AND FURTHER STUDY

The breadth of the study is fairly limited since only two different accounting courses and only two different professors are included in the data. Readers should also recognize that, although the variables used as measures of student and course characteristics exhibit correlations or strong relationships between student test scores, cause and effect cannot be concluded. For example, we cannot conclude that a student's GPA causes their test score on any particular exam. However, the relationship between a student's GPA may help an instructor predict who may need more help in learning information to perform well on a test.

The results show significant decreases in mean test scores as students take the test later in the testing period even when controlling for students' cumulative GPA and Test # within the semester. An estimated marginal means analysis further shows that the upper bound of day 4 is below the lower bound of days 1, 2, and 3, consistent with pairwise comparisons of test score means. The results suggest that information leakage, if any, is not evident in multi-day test scores. The results clearly show that students taking the exam on day 4 are different from students taking the exam on days one through three. The results suggest that an instructor may have an opportunity to further help students taking the exam later in the exam period. Further research on demographics, test preparation, and test taking skills of students taking the exam on day 4 is in order. Perhaps interviews with students can provide a further understanding about student motivation, student test preparation, and student test-taking challenges. Particularly, further research can help instructors learn potential ways to help day four test takers improve their test scores.

Hen and Goroshit (2014) provide some direction for future research on how teachers might find ways to help students. They found that procrastination is related to lower levels of self-regulated learning and academic self-efficacy (Bandura, 1977) and associated with higher levels of anxiety, stress, and illness. They also review and discuss emotional intelligence (EI) and how it may influence a student's ability to assess, regulate, and utilize emotions associated with academic self-efficacy and academic performance including student GPA (see also Haycock, et al., 1998; Wolters, 2003; Zajacova, et al., 2005; Seo, 2008; Klassen et al., 2008; Deniz, et al., 2009). Using the data in the current study, the test starting times showed that day 4 students started the exam on average at 2:51 pm while day one average was 12:39 pm, day 2 average was 1:12 pm, and day 3 average was 1:24 pm. The days of the week showed that most all the tests were taken during weekdays, so weekend test days were not a factor of taking the test later in the day. This data is another indication that procrastination plays a role especially for day 4 test takers. Future research could use standardized tests available to measure students for emotional intelligence, self-efficacy, and motivation, look for direct and indirect relationships to procrastination and academic success. Then instructors might be able to begin to address these related issues to help students be more successful in academic settings.

REFERENCES

Bandura, A. (1977). Selfefficacy: Toward a Unifying Theory of Behavioral Change. Psychological Review, 84, 191215.

Deniz, M., Tras, Z., and Adygan, D. (2009). An Investigation of Academic Procrastination, Locus of Control, and Emotional Intelligence. Educational Sciences: Theory & Practice, 9(2), 623-632.

Haycock, L., McCarthy, P., and Skay, C. (1998). Procrastination in College Students: The Role of Self-Efficacy and Anxiety. Journal of Counseling & Development, 76(3), 317-324.

Hen, M, Goroshit, M. (2014) Academic Procrastination, Emotional Intelligence, Academic Self-Efficacy, and GPA: A Comparison between Students with and without Learning Disabilities. Journal of Learning and Disability, 47(2), 116-124.

Hen, M. and Goroshit, M. (2014). Academic Self-Efficacy, Emotional Intelligence, GPA and Academic Procrastination in Higher Education. Eurasian Journal of Social Sciences, 2(1), 1-10

Klassen, R.., Krawchuk, L., and Rajani, S. (2008). Academic Procrastination of Undergraduates: Low Self-Efficacy to Self-Regulate Predicts Higher Levels of Procrastination. Contemporary Educational Psychology, 33(4), 915–931.

Mouritsen, M., and Davis, J. (2012). Declining Test Score among Introductory Accounting Students: A Comparison of Mean Test Scores in Multi-Day Examination Periods. International Journal of Business and Social Science, 3(15), 1-8.

Muth, J. E. (2006). Basic Statistics and Pharmaceutical Statistical Applications. 256, CRC Press, Boca Rotan, FL.

Reed, S., & Holley, J. (1989). The Effect of Final Examination Scheduling on Student Performance. Issues in Accounting Education, 4(2), 327-344.

Seo, E. (2008). Self-Efficacy as a Mediator in the Relationship Between Self-Oriented Perfectionism and Academic Procrastination. Social Behavior and Personality, 36(6), 753-764.

Wolters, C. (2003). Understanding Procrastination from a Self-Regulated Learning Perspective. Journal of Educational Psychology, 95(1), 179–187.

Zajacova, A., Lynch, S., and Espenshade, T. (2005). Self-Efficacy, Stress and Success in College. Research in Higher Education, 46(6), 677-706.