The positive benefits from the observation that test duration is mostly uncorrelated with student grades

Kevin G. Dunn, kevin.dunn@mcmaster.ca
Department of Chemical Engineering, McMaster University, Hamilton, Canada

Abstract *– New and experienced instructors struggle with setting tests and exams at a suitable level of difficulty, with appropriate questions for the allocated time. Tests that are too short might be thought of as giving students undue advantage. Exams that are too long leave students feeling pressured and anxious, and without time for careful thought to display mastery of the concepts being tested.*

Unlimited time tests are a way to eliminate the effect of anxiety. In this paper we start by reviewing existing work on this topic and explain the data collected in our context. We confirm the literature findings that grades are not inflated by longer durations – if anything, we show there is a slight decrease with longer durations.

Practical applications exist for universities that are facing pressure to shorten exam durations, due to scheduling limitations as class sizes grow. Mainly though, these results will set the mind of new instructors at ease, and validate suspicions of veteran instructors: tests must be of short-enough duration to alleviate time-pressure and anxiety. Building in excess time is required to fairly assess learning outcomes. Students have a higher level of satisfaction knowing they can display their capability fairly, and this comes without undue advantage.

Keywords: test duration, anxiety, stress, new faculty

1. INTRODUCTION

University instructors face many options to assess learning of their course material: class tests, midterms, projects, assignments, quizzes, teamwork, class activities, attendance, presentations, and others. A quick glance at course outlines shows that written midterms and exams are the most heavily weighted elements in courses. Universities schedule and proctor these events for faculty, and unintentionally encourage this form of assessment. Note however, in a few courses, testing under time-pressure might be learning outcome.

Given that so much of a student's grade is determined from tests and exams, it leads to a situation where both instructors and students are nervous about this experience.

Instructors are also anxious about examinations and tests, but for different reasons to their students. When setting an exam, we may have many questions, which when taken collectively, are all aimed at determining whether we are appropriately judging the student's learning:

1. Which type of questions should be used? Long form, short form, multiple choice, fill in the blanks, and many other options exist. Answers to this question are determined by how well the option selected can assess the learning outcomes. Much consideration is also given to how many students are in the class and the time and resources available to grade the subsequent paper. At 15 minutes to grade each examination, this can quickly become an unmanageable burden for moderate class sizes of 100 students. Yet it is common that 50% or more of a student's total grade is determined in this short window of grading time.
2. Instructors wonder what material should be covered in the exam and at what level of difficulty and comprehension each topic should be covered. This decision plays out in courses where students are told that only the latter half of the course is going to be examined, or that only a selection of topics will be examinable. Again, this decision is influenced by the size of the class, and the resources available to grade the handwritten answers.
3. Lastly, instructors are concerned by how many questions should be included, and especially whether the students will finish in the allotted time. Anecdotally, speaking with new instructors, and in this author's experience, it is not unusual that the new instructor sets midterms and exams that are too long in duration. Later on we become better judges of what is reasonable, given the nominal duration of the exam.

Those who have graded exams that are too long can see it in the rushed writing. I have seen direct comments in the exam booklet where the student indicates their

frustration of having run out of time. Emails after an exam indicating they thought it was unfair, or that they could not display their mastery of the concepts being tested have also been received.

Applying material learned in the classroom in real life is mostly, though not always, unconstrained by time. The artificial placement of time constraints in an exam or test is done for the convenience mostly of those administering the exam.

With the above in mind, and after this author's first negative experience with an overly long midterm, we sought to understand the timing effect better by removing time pressure. The procedure followed to collect these data, the subsequent analysis and observations from them are described. First, a review of the existing literature is provided.

2. REVIEW OF EXISTING WORK

It is not uncommon to see media attention drawn to the effect of anxiety during final exams; it is a topic that frequently appears in blog articles and educational mailing lists [1]. Pushes to abolish exams as an assessment method are often seen; for example [2]. Whether that will happen eventually, or not remains to be seen, however, it is indisputable that examinations are a stressful event for those being examined.

Onwuegbuzie and Seaman [3] in their overview of the literature, point out that students prone to higher levels of anxiety cannot live up to their potential in a test. They tend to focus on their anxiety, have task-irrelevant thoughts, and are self-preoccupied, rather than focusing on the examination. Students deemed to have low-anxiety had less interference with their performance, and presumably obtain an advantage because of that.

Hill and Wigfield [4] show results where fifth and sixth graders were either allowed to finish questions attempted (essentially a form of unlimited time testing) or were timed in a way that only two-thirds of all problems could be completed (limited time testing). Participants with high-anxiety scored significantly lower in the timed test, but almost equivalent to the low-anxiety participants when the timing constraint was removed.

Results reported by Orfus [5] agree with the prior two citations, and used a Test Anxiety Inventory to identify high and low-anxiety participants as one factor in their experiments. The second factor used was presence or absence of time pressure. Anxiety had an effect on the outcome (cognitive performance on a math task), while time pressure did not. However, the results do show a strong interaction between anxiety and time pressure, indicating that time-pressure cannot be dismissed as a factor. This result matches the evidence which we have observed (and anecdotally from colleagues) that a subset of students that perform well in the absence of time pressure, such as in class, tutorials and take-home assignments, will sometimes show a sharp drop in tests and final exams, when time pressure is present.

Managing anxiety and time outside the student testing environment was investigated by Case and Gunstone [6]. They consider the experiences of second year chemical engineering students and their perception of time. Interestingly, they found that students perceived they were either in control of time ("*spending time*"; "*saving time*") or that time was beyond their control ("*time caught up with me*"; "*time's not on your side*"), in the context of activities outside of class. As part of their study, they also investigated an unlimited-time class test (midterm), and two regular, limited time class tests. Interviews with students after the unlimited time test mentioned that they felt relaxed, and were able to actually think about the questions, in contrast to their usual experience with time-constrained tests. Similar sentiments are borne out in our results and student quotes, reported below.

Case and Gunstone point out that the instructor's intention of the unlimited time test was to emphasize that it was the understanding (deep learning) that mattered, not the time pressure. The conclusion in their paper is worth quoting directly: "that time-pressured environments can have a deleterious impact on students' approaches to learning and metacognitive development"; and "we may need to radically rethink the prominence of time pressure in our courses" if conceptual, deep-learning approaches are valued.

3. DATA COLLECTION

Experimenting with unlimited time duration tests is only possible in midterms. Final exams, which are centrally scheduled by the university run on a strict timetable and are externally proctored. The mid-semester test on the other hand is completely under this instructor's control.

It is important to point out that written examinations are nonlinear: students can complete or change a prior answer, and they can answer questions out of sequence. Strategy and mental preparation are very much a part of the technique to succeed in such exams. Students have learned and refined this skill over many years, and so the results reported here are relevant, as they are for third year and fourth year courses, implying students should have mastered and internalized a technique to pass exams.

The unlimited midterm durations are nominally set at 2 hours, and the exams start in the evening, around 18:30. Students are told, most times upfront, that this will be an unlimited time exam, where the term "unlimited" implies that they may write the exam for as long as they like, and

are free to leave when they feel they are satisfied with their answers. The longest duration experienced was around 5 hours. A natural constraint is added when scheduling an unlimited time midterm to start in the evening, as students need to return home. Results from 8 such unlimited time tests are reported in Table 1.

Test 1 and Test 7 in table 1 are exceptions that occurred in informing students of the unlimited duration. In test 1, at 45 minutes into the test, it was realized that the test would take too long to complete in the nominal 2 hours. At that point students were told they could write as long as required. Test 7 was a test that nominally should have taken 1 hour to completed, but was scheduled for 2 hours, and all students had to hand in their papers at the 2 hour mark.

The data collected for each test is simply two variables: time to write (measured in minutes) and the student's test score (expressed as a percentage). As students leave the exam we write down the time in hours and minutes. Therefore the time duration for the student is calculated to within an error of approximately 1 minute.

The other variable collected is the student's grade. Grading of midterms is according to a rubric, and is performed by either the instructor or a teaching assistant. It is important to note that each question is graded by only a single person to ensure grading consistency. Typically two teaching assistants and the instructor will grade each student's exam, each assessing about one-third of the total grade. Cross-checking is used for data entry accuracy.

All midterms, except the last one in Table 1, were fully open-book. Students may bring any paper resources with them into the exam: textbooks, notes, prior assignments, prior exams and tests, including solutions. Any calculator is allowed as well, whether programmable or not. Unfortunately electronic textbooks and other electronic devices are not allowed. This testing set up is, from the student's perspective, the most advantageous situation, and matches professional environments as closely as possible (apart from not having access to the Internet).

Test 8 reported in Table 1 had a crowd-sourced formula sheet provided, which the class jointly created two days prior to the test. No aides were permitted in that test, other than a calculator and said formula sheet.

Data from students that receive accommodations are not included, since they write tests in a different venue where the time cannot be observed. These represent about 1 or 2 students per course. Occasionally a student will leave without the time being recorded, and these data points are obviously omitted in the analysis below.

All raw data is available for download at http://yint.org/unlimited-time-tests and the data, as well as R script that was used for generating the plots and data analysis that follows. These may be freely downloaded and used under a Creative Commons Zero license.

Table 1: Data collected from 8 course tests, with N student exams collected. Bootstrapped confidence intervals of the slope coefficients (95% confidence level) and the linear model's standard error are reported.

	Course title	Test date	N	95% confidence interval for slope b_T in equation (1)	Standard error
1	Engineering Economics	October 2012	78	-0.041 to 0.181	13.4
2	Reactor Design	February 2013	80	-0.0821 to 0.0406	10.8
3	Engineering Economics	October 2013	89	-0.1665 to 0.0076	11.2
4	Separation Processes	October 2013	61	-0.2369 to -0.0093 *	13.6
5	Statistics for Engineering	February 2013	79	-0.1663 to -0.0332 *	10.4
6	Process Control	February 2014	101	-0.1882 to -0.0272	15.5
7	Process Control (*limited-time midterm*)	March 2014	97	-0.4641 to 0.0214	17.2
8	Optimization for Chemical Engineers	February 2015	40	-0.2437 to 0.3305	14.3

4. RESULTS ANALYSIS

As mentioned, the reasons for this work were due to a first negative experience with an overly long midterm. It was initially hypothesized that longer writing duration from a student would be correlated with slightly higher grades, for at least two reasons. A persistent student, without the stress of a completion time, might feel they can work through and present their answers in a manner that they believe shows their mastery, and so score a higher grade. Secondly, since all tests were fully open book, consisting mainly of applications of theory, it provides students time to locate and present answers, even those students that might not have studied completely ahead of time.

As reported in the review of prior literature, and confirmed here, the data points when plotting the time taken against the grade achieved shows little relationship. An example of this is shown in Figure 1 for test 2.

A regression model through these data, of the form:

$$g = a + b_T x_T \qquad (1)$$

may be fit to these data, where g is the student's grade. The intercept a is not of any importance in this regression, as it typical in many regression models. The slope coefficient, b_T, is however of interest, and has units of

percent per minute, indicating the change in grade, on average, for each additional minute that is spent writing. The input variable, x_T, is the student's time duration, recorded in minutes. Rather than report a single estimate of b_T, which is prone to noise, and is not really the objective of this study, we show in Table 1 the 95% confidence interval for this slope coefficient.

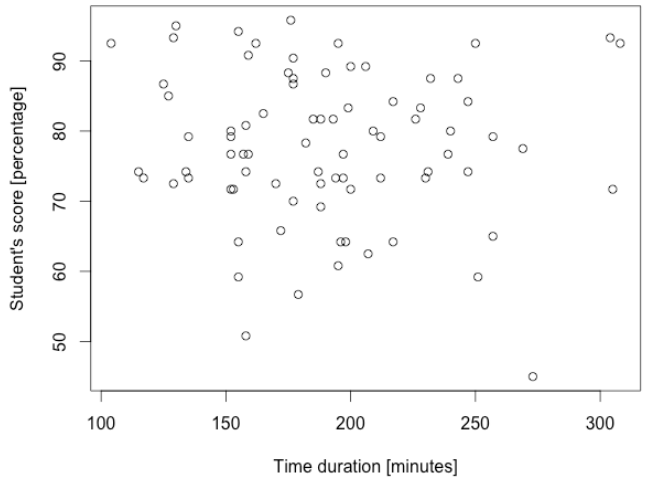

Figure 1: Student's grades, as a function of time duration, for test 2.

Such a confidence interval is far more informative and relevant, as it is the interval within which we expect to obtain the true value of the slope, at the stated level of confidence. The intervals reported are the bootstrap confidence intervals (Efron and Tibshirani [7]), using 10,000 rounds of sampling with random replacement. This implies the least squares model is rebuilt 10,000 times, where N data points are used each time, but the N points are sourced from the original N data points. Some points must then of course be replicated while others are omitted.

This strategy ensures that any outliers in the raw data do not unduly influence the model's slope, since outliers will be omitted (and included) in the 10,000 rebuilds. It is most useful to visually inspect the histogram of the 10,000 repeatedly-estimated slope coefficients, as done in figure 2 for test 2, and contrast it to the raw data for that same test, which is shown in figure 1.

To save space, and quickly compare the tests, it is more useful to report then, as done in Table 1, the points on the left and right histogram tails, that mark 2.5% of the area on the respective tails, which correspond to the dashed lines.

As an example, we interpret the bootstrapped confidence interval for the second test: -0.0821 to 0.0406, which is also visually illustrated in Figure 2. These values indicate the true slope coefficient lies within that bound at the 95% level. Note that the bound contains zero, indicating, the slope is not statistically significant, and that there is no benefit in using the test duration as a predictor of the final test grade. As seen, most of the slope estimates are near zero, and the vertical lines show the 2.5% tail boundaries.

Another example is shown using the 4th test. The histogram is in Figure 3 and the 95% confidence interval of the 10,000 slope estimates is -0.2369 to -0.0093. This indicates there is some, albeit very small, effect on the student's grade due to test duration. In particular, the midpoint is approximately -0.123, which has units of percent per minute. This indicates that each additional minute writing the midterm has a reduction, on average, of 0.123 percent (0.00123 in fractional terms), on the final grade. In other words, each additional hour is associated with, but does not cause, a reduction of about 7.4%, on average, in the student's grade.

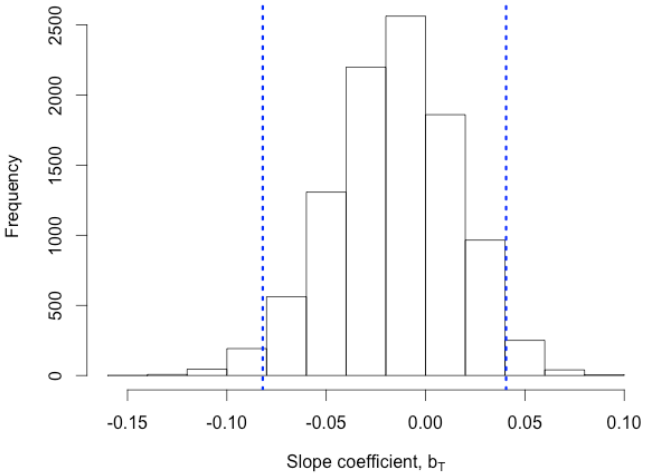

Figure 2: The 10,000 bootstrapped slope coefficients for test 2, and the dashed lines indicate the 95% confidence interval bounds.

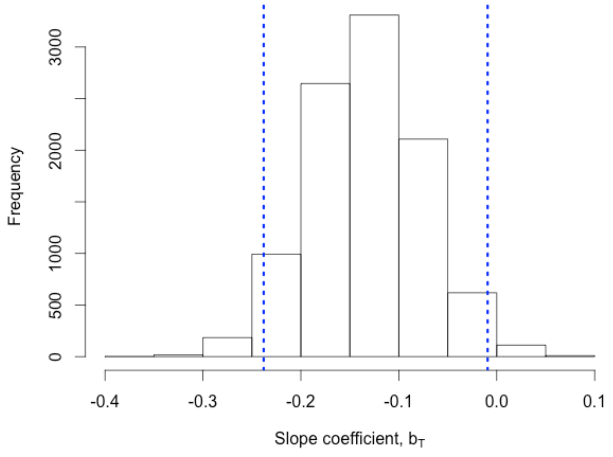

Figure 3: The 10,000 bootstrapped slope coefficients for test 4, and the dashed lines indicate the 95% confidence interval bounds.

It is important to use bootstrapped confidence intervals in this work, to avoid distortion from outliers. Outliers are possible in this data: for example, a student writing for a long time and scoring a very low or high grade could distort the slope. This would be a high-leverage point in the regression model [7, p 268], potentially distorting it. Since our goal is to detect any effect from longer time durations, we want to ensure that outliers will not mislead us with the opposite conclusion. Using these bootstrapped results, in either histogram or confidence interval form, enables us to be convinced of our conclusion.

Test 2 and test 4 were chosen as they had the weakest and the strongest effects respectively; they are two representative extremes, and all others cases lie somewhere in between in terms of their effect. To put this negative effect from test 4 in perspective of a statistic that is widely used, the R^2 value, we have that test 4's model was R^2 =8.9%, while for test 2 it was 0.4%.

The more informative statistic to consider though is the standard error, a measure of the standard deviation of the residuals, if they are normally distributed (which they are). For test 4 it had a value of 13.6, and was at 10.8 for test 2. Both values are comparable, indicating the very wide error spanned by the model residuals. This magnitude of error reaffirms these linear regression models essentially have no predictive value.

To better understand this slight negative effect in test 4, it is worth comparing student grades at the start, middle and end of the test. The first 10 students finishing, the last 10 students finishing, and the remainder in the middle are used to construct the box plot shown in figure 4 for this test. A group size of 10 is used to get a sense of the differences at the extremes. The clear downward trend of the solid median line is observed. Interestingly though, the spread of the whiskers in the first and last box plot is fairly comparable.

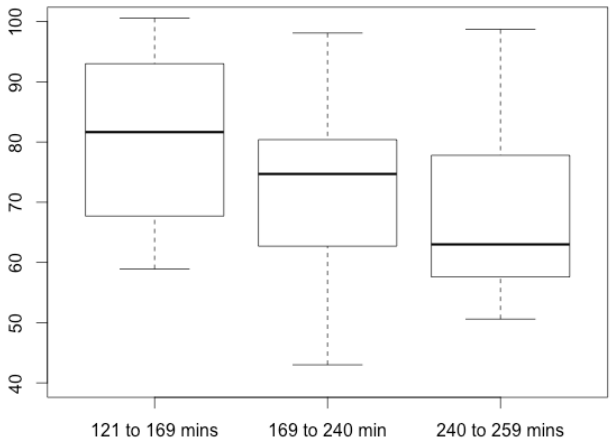

Figure 4: Box plots of the grades for the first 10, middle group, and the last 10 students in test 4.

A similar box plot sequence for test 2 is shown in figure 5, the test which was most neutral. This shows a very slight decline in the solid median lines, and the last 10 students finishing have a wide variety in their final grade as indicated by the spread of the last box plot in the sequence. This validates the raw data in figure 1 and the histogram of slope coefficients in figure 2.

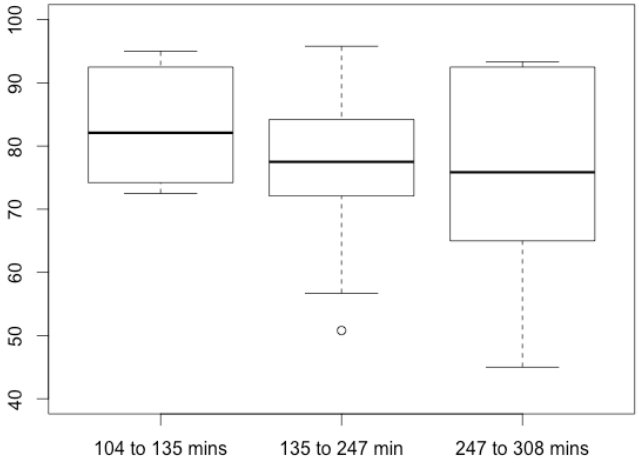

Figure 5: Box plots of the grades for the first 10, middle group, and the last 10 students in test 2.

4. RESULTS INTERPRETATION

We emphasize that the analysis of the data is entirely correlation-based. No cause and effect mechanism is being implied. Correlation does not imply cause and effect, but correlation is a prerequisite for cause and effect. We may then postulate two possible causal mechanisms below, which we have termed strategizing and self-doubt, and propose ways in which each mechanism might be investigated in the future.

More reliable cause-and-effect data would require active experimentation where students are randomly divided into groups and each group would be limited to certain times (e.g. 120 minutes, 140 minutes, 160 minutes, *etc*) as was done in the study by Wright [9]. This would be almost impossible to administer to a consenting single cohort of students in a university class.

So these results only show that that longer times spent in the test are correlated with negligible to small decreases in the average grade.

1. Strategizing

When students know the exam is open book and of unlimited time, they might be less conscientious in studying before the exam. They choose to use the unlimited time available in the midterm to learn the material, and answer the questions. To do so would result

in a longer exam duration, and would not necessarily result in scoring a higher grade, as the help they need to fully understand the material at that moment is not available to them. They would certainly achieve a higher grade than in the time-limited case. This is likely the reason students expressed satisfaction with the unlimited time test option.

This strategizing hypothesis is postulated for two reasons. The first 3 tests that showed no statistical effect were on a new cohort of students that had never experienced this lack of time pressure before. Students writing tests 4, 5 and 6 however had one prior experience with this form of testing, and so could potentially strategize as described above.

The second reason for proposing this mechanism is based on student responses. These data were given to the author's students in his class, Statistics for Engineering, to analyze as part of an assignment. The students were asked to answer the question: "What advice would you give to students based on these results?" and "What result(s) do you learn from these data that is(are) useful for course instructors to know?" One student wrote: "*The best way to treat the test is just like any other midterm. Study beforehand and understand the concepts in order to do well.*" Another student wrote: "*Extra time may not aid the student if they do not have a solid understanding of the material.*" Verbal discussions with students indicate they were counting on using the unlimited test time to both study and answer the questions. The results presented above bear out that this strategy is not a beneficial one.

To confirm the strategizing effect, a self-reported survey could be taken after the midterm which asks students to report the approximate time spent preparing for this exam. Whether the survey would be honestly and accurately answered is another matter. This survey could be extended to understand the time pressure effect further.

2. Self-doubt

The second possible mechanism for the slight drop in grades is that students might doubt their prior answers, scratch out or erase them, and rewrite an alternative. This would require the student to be present in the exam room for a longer time. This rewriting might not necessarily improve their grade if they are rewriting an answer incorrectly.

The reason for this proposed mechanism is based on the observation of erased and corrected answers in the booklet. However there is uncertainty whether the level of corrections is any higher to a regular time-constrained test. This is a weaker reason, and likely coexists with the prior mechanism, as students which are strategizing are almost certainly attempting and reattempting questions.

One of the student responses in the assignments was: "*I would say this is a really good way to actually find out what a student actually know because some people are not as fast as others and they start to stress out towards the end of the exam as times is about to run out. By having unlimited time tests, it creates more real world situation, as in real world there usually would not be as much time induced pressure to complete a task.*"

Studying this mechanism is difficult, but could be done by closely examining test scripts for corrections and changes. However, as tests go increasingly online, it is possible to track corrections and changes made far more accurately, whether from initially correct to incorrect, or *vice versa*. Tracking periods of activity and inactivity is also possible in online tests, and might be used to investigate the observation that weaker students often stay in the test venue without writing or working on the test.

5. BENEFITS AND IMPLICATIONS

The main benefit of this form of unlimited time testing, given the context of the anxiety and time-pressure effects discussed in section 2, is that we can fairly assess students, no matter their level of anxiety, when the time-pressure is removed. The converse is not true for all students: **if there is not enough time to complete the test, we likely cannot obtain a fair assessment of the student**, since time-pressure induced anxiety will play a negative role for a subset of the students.

Having a longer time to write their test gives students the ability to reflect on their written answers, erase and rewrite them, or modify them to improve their answer. If the goal of testing is to obtain a fair assessment of the student's knowledge, and not how fast they transcribe and arrange their thoughts, then an unlimited time test is warranted.

In cases where the cost of a longer test duration is fairly marginal, then this would be advised, together with clear communication to students for the intention of this extra time. The author has also found it valuable to provide the raw data to students to interpret, as a way for students to self-confirm the effects described in this paper.

Where testing time cannot be lengthened, it places even greater onus on the instructor to carefully craft questions that test the student's complete understanding. Using only a limited number of questions that can be approximately completed in 50 to 60% of the allotted time available should suffice. The recommendation is to add a sentence on the exam to the effect that it can be completed in a short time, however excess time is built-in to reduce time pressure. This would hopefully assuage concerns of students prone to test-induced anxiety.

To conclude, students seldom commented on the unlimited time midterm in their course evaluations. No negative comments were received. Some positive comments were noted in one evaluation: "*Having open book and no time limit on the midterm relieved all the stress and allowed me to really convey my knowledge and ability a lot more effectively. Please don't ever stop doing this*."

Further work on this topic was proposed by one student who commented in an assignment that: "*Studying time, a students* [sic] *passion about the material, intuitive understanding can all contribute to the mark on the test, so one should not conclude that writing time is the sole factor of* [the] *mark obtained. Further studies should be done which encompass all variables in addition to* [time]".

References

[1] Society for Teaching and Learning in Higher Education Listserv, http://www.stlhe.ca/listserv/

[2] Joseph Brean "The death of the exam: Canada is at the leading edge of killing the dreaded annual 'final' for good" National Post, 02 April 2015.

[3] Anthony J. Onwuegbuzie and Michael A. Seaman, "The effect of time constraints and statistics test anxiety on test performance in a statistics course", *The Journal of Experimental Education*, vol. 63, no. 2, pp. 115-124, 1995. http://www.jstor.org/stable/20152442

[4] Kennedy T. Hill; Allan Wigfield "Test anxiety: A major educational problem and what can be done about it", *The Elementary School Journal*, vol. 85, no. 1, pp. 105-126, 1984, http://eric.ed.gov/?id=EJ307229

[5] Shauna Orfus, "The effect test anxiety and time pressure on performance," *The Huron University College Journal of Learning and Motivation*, vol. 46, no. 1, 2008. http://ir.lib.uwo.ca/hucjlm/vol46/iss1/7

[6] Jennifer Case and Richard Gunstone, "Going deeper than deep and surface approaches: A study of students' perceptions of time", *Teaching in Higher Education*, vol. 8, no. 1, 2003 DOI: 10.1080/1356251032000052320

[7] Efron, B and Tibshirani, R. *An Introduction to the Bootstrap*. Boca Raton, FL, Chapman & Hall, 1993.

[8] John Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, Sage, 1997.

[9] Ted Wright, "The effects of increased time-limits on a college-level achievement test", (Report No. 84-12). Miami-Dade Community College, FL Office of Institutional Research, http://eric.ed.gov/?id=ED267867