3-2017

Short vs. long: cognitive load, retention and changing class structures

Brandon Sheridan
Elon University

Ben O. Smith
University of Nebraska at Omaha, bosmith@unomaha.edu

Erin G. Pleggenkuhle-Miles
University of Nebraska at Omaha, erinpmiles@unomaha.edu

Short vs. Long*

Cognitive Load, Retention and Changing Class Structures

Brandon Sheridan[1], Ben Smith[2,†], and Erin Pleggenkuhle-Miles[3]

[1]Department of Economics, Elon University, bsheridan4@elon.edu, 2075 Campus Box, Elon, NC 27244
[2]College of Business Administration, University of Nebraska at Omaha, bosmith@unomaha.edu, 6708 Pine Street, Omaha, NE 68182
[3]College of Business Administration, University of Nebraska at Omaha, erinpmiles@unomaha.edu, 6708 Pine Street, Omaha, NE 68182

2017

Keywords: Cognitive Load, Class Structure, Retention, Student Performance, Class Length

*The authors thank Jeff Bredthauer, Bill Goffe, James Tierney, Jamie Wagner and Dustin White for their comments on early drafts of this paper. We further thank Colin Green and the anonymous referee for their helpful comments.
†Corresponding author.

Abstract

University class structure is changing. To accommodate working students, programmes are increasing their offerings of long night classes – some lasting as long as six hours. While these long classes may be more convenient for students, they have unintended consequences as a result of cognitive load (Van Merrienboer and Sweller, 2005). Using a panel of 124 students (372 observations) and a differencing approach that controls for student characteristics, we show that student exam performance decreases by approximately one-half letter grade on content taught in the second half of a long class (significant at the 5% level).

1 Introduction and Review of Literature

Starting in the 1990s, a number of U.S. universities began offering part-time programmes. These part-time programmes allowed [potential] students who worked during the day to take classes at night, and were quick to gain popularity. Using the most popular master's degree programme as an example (MBA), half of all programmes are part-time and 50% of those have experienced applicant growth since 2011 – double that of the full-time programmes (Rafferty, 2013).

This growth of U.S. part-time programmes has resulted in an increase in multi-hour class offerings – usually scheduled at night. For the convenience of both students and administrators, these classes often meet a single time per week; requiring each class session to be at least twice as long as their daytime counterparts. If the term has been shortened as well, as is common in MBA programmes, the length of each class session is increased further. This practice seems to be most pronounced on urban campuses.

Amongst the 90 members of the Coalition of Urban and Metropolitan Universities (CUMU) 84 offered publicly available schedules[1]. Of the 76 members with MBA programmes, 91% offer classes where each session lasts at least two hours and 30% offer class sessions that last three hours or more. 52% offer undergraduate economics classes with sessions in excess of two hours; 15% offer undergraduate classes with sessions that last three hours or more.

While one might assume this practice is limited to regional universities, the evidence suggests otherwise. For instance, the economics departments of New York University, Georgetown, and American University offer applied masters degree programmes that can be completed by part-time working students at night; these prestigious programmes are not an anomaly, but a reflection of a trend on urban campuses. Of the 54 institutions in the Coalition that offered an MA or MS in economics, 79% offered classes with sessions of two hours or more and 19% of institutions offered

[1] Six members of CUMU require users to log in with an institution account to see course schedules. In all cases, we collected class length from the Fall 2016 schedule.

class sessions lasting three hours or more.

This growth in long night classes unfortunately runs counter to the cognitive load literature – a field concerned with implications of working memory. Numerous studies (Karpicke, 2012; Karpicke and Grimaldi, 2012; Karpicke and Smith, 2012; Roediger and Butler, 2011) have found that retention is a function of recalling information (retrieval) repeatedly. Both the spacing and repeated nature of the process can impact retention with studies showing the importance of retrieval events occurring over spaced intervals (Cepeda et al., 2006, 2008).

Mental resources, such as working memory and processing, are limiting factors when a student is learning a new topic. Building on the work of Baddeley and Hitch (1974), studies have shown that a trade-off exists between mental processing and storage whereby performance decreases when the subject is attending to more than one task – described in the literature as concurrent memory load (Anderson et al., 1996; Case et al., 1982; Conway and Engle, 1994). This condition is further exacerbated if the subject is 'overwhelmed' by the difficulty of understanding the new information. Anderson et al. (1996) has shown that as mental processing increases in difficulty, information can be lost from short-term memory.

With the development of cognitive load theory (Sweller, 1988), instructors have been given tools to better allocate their students' cognitive resources during learning tasks. From an instructor's point of view, one can allocate a student's mental resources to cognitive tasks that are helpful (germane load) while attempting to reduce unnecessary load (extraneous load). Much of the cognitive load literature focuses on the reduction of extraneous load (Cierniak et al., 2009; Sweller et al., 1998; Ward and Sweller, 1990). For example, there are a number of studies that look at different mechanisms instructors can use to reduce in-class cognitive load (Sweller, 1994; Sweller et al., 1998), such as the use of pre-lecture resources (Seery and Donnelly, 2012). Rather than addressing components of pedagogy under the control of the instructor, our study focuses on the impact of longer classes on the available mental resources of the student.

With longer classes (such as those that meet once per week), resource availability can be impacted by path dependence and mental fatigue. Many times in a course an instructor will introduce a concept and then build on that concept in a later section. For the student, using this new concept requires a great deal of working memory, but with sufficient practice many tasks can become automated (Kotovsky et al., 1985; Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977). For instance, suppose an instructor is teaching a concept that requires the student to graph points of data. The instructor might first dedicate some class time to the graphing technique itself then move on to the economics problem in the next hour of class. In short classes (such as those that meet multiple times per week), the student would have the ability – and often the requirement – to study (practice) the graphing technique multiple times before the economics concept is taught, reducing the amount of necessary working memory (Kalyuga et al., 2001). This underscores a fundamental issue with longer class sessions. In long classes, the student must study both the graphing technique and the path dependent economic concept at the same time as both concepts were taught in the same class period. This requires the student to work harder, as he/she engages their working memory to learn each new topic that builds upon the previous.

While the issue of concentrated learning has been studied in the context of K-12 (primary and secondary education) and medical education, no previous study shows the impact of long classes in a university classroom setting. The work of Anderson and Walker (2015), Hewitt and Denny (2011), and Yarbrough and Gilman (2006) show little to no evidence of a detrimental learning impact due to K-12 districts adopting four day school weeks. However, K-12 is very different from college; studying (practice) often occurs in the classroom in a K-12 environment while it generally occurs at home at the university level. In contrast, Raman et al. (2010) shows that long-term learning of medical residents is improved when material is presented over multiple sessions and Pope and Fillmore (2015) finds that students who take multiple Advanced Placement exams score better when they have more time between exams to prepare.

With the prevalence of curriculum and programme redesigns taking place in urban universities, this salient topic is of possible interest to administrators, department chairs and instructors – especially those teaching long classes or serving on programme committees. Our findings also might encourage universities to investigate solutions where long classes are shortened by some administrative mechanism. This manuscript proceeds as follows. In section 2, we discuss our quasi-experimental method. In section 3 we discuss the results from two different courses (subsections 3.1 and 3.2). In subsection 3.3, we pool the entirety of the dataset (124 individuals with a total of 372 observations) and show our results are significant at the 5% level. This is followed by robustness checks of our result (subsection 3.4). Finally, in section 4, we conclude.

2 Method

There is an inherent selection problem in comparing two sections of the same course. Working students, who are often older, are more likely to take the longer night classes. Further, course load, work and family obligations are not likely to be independent of students' class selection. In this manuscript, we use a differencing method where we use each students' performance on half of the course content as the control. This approach allows us to control for the aforementioned selection issues without the need for a long list of demographic characteristics.

During this experiment, one of the authors taught at 'College A.' College A is a small private liberal arts institution offering both 'short' daytime and 'long' evening classes[2]. Daytime classes are offered either twice or three times a week ('short' classes). Evening courses have class sessions that run two or three times as long as their daytime equivalents, but only meet once per week ('long' classes). As College A operates on ten week terms, each long class is three and a half hours in length.

[2]Our study was conducted at College A for experimental design reasons. For our experiment to be valid we needed two sections of the same course taught by the same instructor in two different formats: one during the day and one

The economics faculty at College A routinely teach three or four classes. However, some classes may be multiple sections of the same course offered at different times of day. In the Fall of 2015 and the Spring of 2016, the author taught two short/long class pairs: Money and Banking and Intermediate Macroeconomics. These two courses are quite different from one another. The former is an elective for economics majors, requires many relatively straightforward mathematical calculations, and (anecdotally) is considered relatively easy. The latter course is required, contains more complex mathematical calculations, incorporates many graphical elements, and (again, anecdotally) is considered to be more difficult by the student body.

Both sections of each course received the same exams and assignments. Furthermore, they were taught in a nearly identical manner. As a way of reducing workload, the instructor kept the classes in sync such that the i'th hour of class was the same regardless of format. Put another way, content taught on the first day of the week in the short daytime class would be taught in the first half of the long nighttime's class period. Equivalently, content taught on the last day of the week of the short daytime class would be taught in the second half of the long nighttime class lecture. Throughout this manuscript, we will use the term 'first half of the week' to refer to the first 1.75 hours of class time in a week – half of the 3.5 weekly total – regardless of whether the class is delivered over multiple sessions per week (daytime) or a long single session (nighttime). We will similarly define the term 'second half of the week' as the last 1.75 hours of class time in a week.

Although approximately half of each class was of the traditional lecture format, the instructor incorporated numerous other pedagogical techniques. For example, the instructor frequently used videos, podcasts, and websites (especially for primary source data of relevant information such as the unemployment rate), and often followed these up with a class discussion. Additionally, cooperative learning exercises were utilised during many class periods to provide practice for complex

at night. For this reason, we could not conduct our experiment using a multi-hour MBA or similar course. While our results are relevant to part-time master's degree programmes, most universities do not offer both a short and long version of such courses.

topics. As such, estimates of the impact of the longer class period should be attenuated by these active learning strategies.

Throughout the semester, the instructor tracked what content was taught in which half of the week. This content was then mapped to exam questions – which were identical in both classes. Using this information, we can control for student abilities and selection issues. Our dependent variable is the difference between performance on content taught in the second half of the week (back) and performance on content taught in the first half of the week (front). We use the following empirical model to aid in our investigation.

$$S_{ieB} - S_{ieF} = \beta L + \gamma_e + \varepsilon_{ie} \tag{1}$$

where S_{ieB} is student i's percent correct score [0-1] on exam e on content taught in the second half of any given week. S_{ieF} is student i's percent correct score on exam e on content taught in the first half of any given week. γ_e is exam e's fixed effects and ε_{ie} is the error term. As L takes a value of one for any student in the long (night) class, β is our coefficient of interest.

Our identification strategy is akin to a difference-in-difference method where we control for individual ability through one difference and for dissimilar content difficulty through another[3]. Our dependent variable is the difference in performance between content taught in the second half of the week versus the first half of the week for each student/exam observation. As each student's performance on the second half of the material is differenced from their own performance on the first half of the week material, all time-invariant characteristics, such as student quality, are removed. In our study, time-invariant factors are characteristics common to both the first and second half of the week of a given exam. Time-variant factors, such as content difficulty, differ from the first to the second half of the week of a given exam. Another advantage of this approach

[3] A general explanation of methodology can be found in Angrist and Pischke (2008), pp. 227–243. For a practical example see Card (1992).

is it allows the student control to vary by exam. For instance, if a student did poorly on the first exam but then implemented better study habits for exams two and three, this method would control for this change, while standard student fixed effects would not.

Each difference in performance (second half of the week versus first half of the week) can be explained by a few factors: differential difficulty of the material, time-variant idiosyncratic factors and a difference in the length of the class. To control for the difference in material difficulty, we have a difference in difficulty fixed effect (γ_e) for each of the exams. These difficulty fixed effects are identified by the short (daytime) class students as they have been exposed to the difference in difficulty of the two sections of the exam but have not been exposed to the structural difference (i.e. the class length). The remaining variation can be explained by the length of the class and time-variant idiosyncratic factors. Assuming the time-variant idiosyncratic factors and class structure are uncorrelated, then our estimation of β is accurate.

While an unobserved time-variant idiosyncratic factor shared by the treatment group is always a concern, it seems unlikely in this study. Our treatment (the structure of the class) was administered six times in the form of an exam (three exams in Money and Banking and three exams in Intermediate Macroeconomics). With each of those exams, the treatment was the same: class length. But, the time-variant component (exam content) differed. Thus, the primary potential source of a common time-variant idiosyncratic factor was different with each treatment. Therefore, it is unlikely that this common time-variant idiosyncratic factor would exist with each administration of the treatment; if a common time-variant idiosyncratic factor existed it would likely result in differences in the estimated impact of the class structure across treatment administrations. However, we show in section 3.4 that our point estimates are consistent across students and exams.

In the following section we discuss the results of this quasi-experiment for both Money and Banking (section 3.1) and Intermediate Macroeconomics (section 3.2). We then pool the data for our main result (section 3.3) – which is significant at the 5% level. Section 3.4 demonstrates that our result is driven by the entirety of the dataset and not a common subgroup.

3 Results

3.1 Money and Banking

The two paired sections of Money and Banking were given three exams ranging from 23 to 26 multiple choice questions. Each of the 29 short class daytime students and 35 long class nighttime students took every exam. Scores are coded from zero to one representing the percent of questions the student answered correct. Table 1 provides the exam summary statistics for both sections of the course.

[Table 1 about here.]

In both sections of Money and Banking, exams accounted for 75% of the final grade (25% each). Traditionally, academic grading in the United States maps the overall weighted percentage of exams and other assignments to a letter grade by 10 percentage point increments. That is, final weighted percentages between 90-100% would receive an A range grade (A+, A, A-), those between 80-89.9% would receive a B range grade (B+, B, B-) and so on. Overall, the median final grade in Money and Banking was a B-, the maximum grade was an A and the minimum grade was an F.

Roughly 43% of exam questions were taught in the first half of the week with the remainder being taught in the second half of the week. However, while there were slightly more exam questions pertaining to content taught in the second half of the week, both sections received the same exam and the content was taught during the same portion of the week.

[Figure 1 about here.]

As a first test, we examine the mean difference of exam performance based on which part of the week the material was taught. As shown by Figure 1, the short (daytime) class performed about 0.052 points better on material taught in the second half of the week than the first ($\bar{S}_{B_S} - \bar{S}_{F_S}$). However, the long (nighttime) class performed only 0.020 points better on material taught in the second half of the week ($\bar{S}_{B_L} - \bar{S}_{F_L}$). This suggests that the impact of the treatment is approximately -0.031 – what we refer to in our table as the mean difference of differences ($(\bar{S}_{B_L} - \bar{S}_{F_L}) - (\bar{S}_{B_S} - \bar{S}_{F_S})$). Please note, in calculating the difference of differences presented in Figures 1, 2 and 3, the authors do not round the mean values to three digits first. Therefore, our values may differ slightly from the reader's calculation based on the values presented in the table.

Additionally, we provide a t-test (Welch, 1947) of the mean short class difference ($\bar{S}_{B_S} - \bar{S}_{F_S}$) versus the long class difference ($\bar{S}_{B_L} - \bar{S}_{F_L}$). Due to the independence assumption of a t-test, we grouped and averaged the differences for each student before performing the t-test. This has no impact on the mean value, but is a more conservative method of testing significance. While we do not consider this t-test our primary method, it hints at the results we will obtain when we move to a formal regression model.

[Table 2 about here.]

Using ordinary least squares (OLS) and our empirical specification in section 2, we are able to provide the regression results shown in Table 2. With a difference in performance on the exam as the dependent variable ($S_{ieB} - S_{ieF}$), we show that the long (night) class's difference in performance is about three percentage points worse than the short class's difference. Using White's Lagrange Multiplier test (White, 1980) and a Shapiro-Wilk test (Shapiro and Wilk, 1965), we cannot reject the null hypotheses of homoscedastic and normally distributed errors. Nonetheless, we report cluster-robust standard errors (Williams, 2000) as there is a potential lack of independence across the three observations per student. Further, a bootstrap procedure[4]results in a similar p-value to our cluster-robust approach (0.209). While these results are not statistically significant by themselves, we ask the reader to take note of the point estimate. In the upcoming subsections we will find nearly identical results in the Intermediate Macroeconomics and the pooled model – both of which are statistically significant.

3.2 Intermediate Macroeconomics

Similar to Money and Banking, the two paired Intermediate Macroeconomics sections were given three exams. All three exams were taken by all 29 short (day) class and 31 long (night) class students. The first two exams consisted of 25 questions each; 12 of which related to content taught in the first half of a given week (48% of each exam). The final exam contained 25 questions, 20 of which were taught in the same hour of class for both sections – the remaining five questions were excluded from our analysis due to the time mismatch. Of those 20 questions, 10 related to content taught in the first half of the week. Table 3 shows the exam summary statistics for both sections of Intermediate Macroeconomics.

[Table 3 about here.]

Like Money and Banking, exams accounted for 75% of the Intermediate Macroeconomics final grade (25% each). Similarly, final grades were assigned using the same academic norms described for Money and Banking. Overall, the median final grade in Intermediate Macroeconomics was a C+, with a maximum grade of an A and a minimum grade of an F.

[4] Throughout this manuscript, we use a bootstrapping procedure such that we randomly select individual students, not rows, in the panel with replacement. Then using the randomly selected individuals (which may appear more than once and some may not appear at all), we select the corresponding rows – some of which may be duplicated if a particular student was randomly selected multiple times. This block bootstrap (Künsch, 1989) is an appropriate method for panel data where errors could be correlated to the individuals. However, we also performed a pair-wise bootstrap where we resampled the rows. These procedures produced nearly identical p-values for our dataset and in all cases the pair-wise p-value was smaller than the reported value produced by resampling the individual students. We run all bootstrap procedures at 10,000 repetitions.

In an identical fashion to Money and Banking, both sections received exactly the same exam. However, unlike Money and Banking where the short class appeared to be a slightly stronger group, the Macroeconomic long (night) class seems to dominate the short (day) section. Nonetheless, our method should control for these differences in characteristics.

[Figure 2 about here.]

Figure 2 shows the performance of the two Intermediate Macroeconomics classes based on when the content was taught. Like Money and Banking we will compare the means of the sections by when the content was taught as a first pass. The short (daytime) section performed about 0.001 points better on material taught in the second half of the week than the first ($\bar{S}_{B_S} - \bar{S}_{F_S}$). Nonetheless, the long (nighttime) class performed 0.048 points worse on the second half of the week material than they did on the first half ($\bar{S}_{B_L} - \bar{S}_{F_L}$). This would imply that the impact of the long class (mean difference of differences) is about -0.049 ($(\bar{S}_{B_L} - \bar{S}_{F_L}) - (\bar{S}_{B_S} - \bar{S}_{F_S})$).

As in the previous section, we provide a simple t-test (Welch, 1947) of the short class mean difference ($\bar{S}_{B_S} - \bar{S}_{F_S}$) from the long class mean difference ($\bar{S}_{B_L} - \bar{S}_{F_L}$). Due to the independence assumption of a t-test, we group and then average each student's three observations before performing this operation. While we do not consider this t-test to be our primary way of determining significance, it suggests that we will find our results significant at the 10% level when we move to a formal regression model.

[Table 4 about here.]

Using OLS and our specification in section 2, we provide the regression results in Table 4. With a difference in performance on the two halves of the exam as the dependent variable, we show that the long (night) class's difference in performance on second half of the week material is about four and a half percentage points worse than the short (day) class; this result is significant at the

10% level. Using White's Lagrange Multiplier test (White, 1980) we can reject the null hypothesis of homoscedastic errors. Therefore, due to the potential lack of independence within the panel, we have reported the cluster-robust (and heteroscedasticity-consistent) standard errors (Williams, 2000). Using a Shapiro-Wilk test (Shapiro and Wilk, 1965), we cannot reject the null hypotheses of normally distributed errors. However, using the same bootstrap procedure as before, the long dummy p-value is 0.063.

3.3 Pooled Model

Finally, we pool the datasets from the Money and Banking and Intermediate Macroeconomics classes, which yields a total of 124 students and 372 observations. We examine the pooled data in a similar manner to Figures 1 and 2, then move to a full regression model.

[Figure 3 about here.]

In Figure 3 we present the adjusted differences of the short and long classes. Due to the dissimilar difference in difficulty of Money and Banking and Intermediate Macroeconomics, we have re-centred the distributions at zero (for the purposes of Figure 3 only; we have not modified the data for our regressions). That is, we have added 0.051 to each Money and Banking student's performance difference by part of the week. Similarly, we have added 0.001 to the performance differences in the Intermediate Macroeconomics courses. This moves the distributions such that both the short (daytime) Money and Banking and Intermediate Macroeconomics are centred on zero but maintain their original distance to their long (nighttime) counterparts. This allows us to treat both short classes as a single set of observations and similarly treat both long classes as another set of observations.

As in sections 3.1 and 3.2, we perform a t-test (Welch, 1947) as a first step in our analysis. Using the data in Figure 3, we group our observations by student and calculate the mean. We then compare the short class distribution mean to the long class distribution mean – the difference is significant at the 5% level.

In Table 5, we present the results of our OLS regression. Like the results in sections 3.1 and 3.2, the dependent variable is the difference in performance of student i on exam e. However, we create separate exam fixed effects for each of the two courses. In this specification, γ_{eMB} is the Money and Banking exam e fixed effect and γ_{eIM} is the Intermediate Macroeconomics exam e fixed effect.

[Table 5 about here.]

Similar to the individual class results, the point estimate for β is approximately -4%. This result is significant at the 5% level. Using White's Lagrange Multiplier test (White, 1980), we reject the null hypothesis of homoscedastic errors. Therefore, we report cluster-robust (and heteroscedasticity-consistent) standard errors (Williams, 2000). Like the individual course results, using a Shapiro-Wilk test (Shapiro and Wilk, 1965), we cannot reject the null hypothesis of normally distributed errors. However, the p-value of β using a bootstrap reveals a p-value of 0.033.

3.4 Robustness of Results

In this subsection, we analyse portions of our observations to determine if our results are anomalous. Two possibilities exist that we wish to test. First, our main pooled result (section 3.3) could be driven by a single exam during the semester. For instance, perhaps students are more susceptible to path dependence later in the semester. Second, we will determine if our results depend on overall exam performance.

To determine if our point estimates vary throughout the semester, we have augmented our original empirical model with two additional variables: L_{FE} and L_{SE}. L_{FE} has a value of one if the student is both in the long (night) class and this is the first exam of the semester. Similarly, L_{SE} has a value of one if the student is both in the long class and this is the second exam of the semester.

As L continues to take a value of one for all exams in the semester, the third exam is our baseline. Therefore, the point estimates generated from L_{FE} and L_{SE} will show the difference in impact of the long class for the first and second exam in comparison to the third. We can restate our regression equation as follows:

$$S_{ieB} - S_{ieF} = \beta L + \beta_{FE} L_{FE} + \beta_{SE} L_{SE} + \gamma_e + \varepsilon_{ie} \tag{2}$$

Using equation 3, we obtain the results in Table 6. Our variable of interest (β) is very close to our point estimate in section 3.3. In comparison to the third exam (the baseline), long class students performed slightly better on second half of the week material on the first exam and slightly worse on the second exam. Nonetheless, both point estimates are extremely small indicating that exam order is not substantively impacting our results.

[Table 6 about here.]

A second possibility is that our point estimates are related to performance on the exam. It is well known that better students have better metacognition (Coutinho, 2008). Therefore, it is possible that better long (night) students are aware that their understanding of material in the back half of the week is slightly worse and adjust their studying accordingly.

For each exam of the four sections, we split the observations into two groups: those with an overall exam score above the mean for a given exam and those with an overall exam score below the mean. We then coded all below the mean observations with a value of one and stored it as the variable B (zero otherwise). This method of splitting the observations ensured that each of the exams were roughly equally weighted in the two subgroups. We then performed a OLS regression using the following modified empirical model:

$$S_{ieB} - S_{ieF} = \beta L + \beta_{LB} L \times B + \beta_B B + \gamma_e + \varepsilon_{ie} \tag{3}$$

The results in Table 7 suggest that the difference in the dependent variable in the long (night) class is driven by both the low performing exams and high performing exams. If our results were driven by the low performing students, we would expect to see β_{LB} capturing most of the effect and β moving towards zero. Conversely, if our results were driven by the high performing students β would be more negative and β_{LB} would be an offsetting positive number. However, we find β_{LB} to be near zero. Splitting the data using the median or splitting the students (as opposed to exams) with their average exam score produces similar conclusions.

[Table 7 about here.]

4 Conclusion

In this paper we show that there is approximately a three to four percentage point decrease in performance in the second half of a long class. Our results are statistically significant for both the pooled sample and the Intermediate Macroeconomics class. Further, the point estimates are consistent across both courses we tested and the individual exams. Therefore, our results demonstrate a consistent effect not driven by a single exam or class anomaly. Our results are strengthened by the differences between the two courses taught, since each caters to a different subset of the student body.

The observed performance difference is not the result of different choices made by the long (nighttime) versus the short (daytime) student, but rather is a difference in mental capacity due to the current cognitive load of the student. Therefore, we expect our results to be generalisable to other courses and subjects, but more study is needed. While the two courses in this study are very different, they are both economics courses – a field with a high degree of concept path dependance. Courses that are more focused on memorisation may be more immune to this phenomenon. Even if all courses are negatively impacted by extended class sessions, the point estimates are unlikely to be the same in a different field. As there are administrative advantages to longer night classes, the precise 'cost' of longer classes would be of interest to administrators, department chairs, instructors and programme committees.

Despite these limitations, our results suggest the length of class sessions should be considered when redesigning programmes. In some cases, longer classes are unavoidable. Therefore, instructors teaching longer classes should mitigate the structural difference as much as possible and prepare their students to study more than their daytime counterparts.

References

Anderson, JR, LM Reder, and C Lebiere. 1996. "Working memory: Activation limitations on retrieval". *Cognitive Psychology* 30(3): 221–256.

Anderson, MD and MB Walker. 2015. "Does shortening the school week impact student performance? Evidence from the four-day school week". *Education Finance and Policy*.

Angrist, JD and J Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Baddeley, AD and G Hitch. 1974. "Working memory". *The Psychology of Learning and Motivation 8*: 47–89.

Card, D. 1992. "Using regional variation in wages to measure the effects of the federal minimum wage". *Industrial & Labor Relations Review* 46(1): 22–37.

Case, R, DM Kurland, and J Goldberg. 1982. "Operational efficiency and the growth of short-term memory span". *Journal of Experimental Child Psychology* 33(3): 386–404.

Cepeda, NJ, H Pashler, E Vul, JT Wixted, and D Rohrer. 2006. “Distributed practice in verbal recall tasks: A review and quantitative synthesis.” *Psychological Bulletin* 132(3): 354.

Cepeda, NJ, E Vul, D Rohrer, JT Wixted, and H Pashler. 2008. “Spacing effects in learning a temporal ridgeline of optimal retention”. *Psychological Science* 19(11): 1095–1102.

Cierniak, G, K Scheiter, and P Gerjets. 2009. “Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?” *Computers in Human Behavior* 25(2): 315–324.

Conway, A and R Engle. 1994. “Working memory and retrieval: a resource-dependent inhibition model.” *Journal of Experimental Psychology: General* 123(4): 354.

Coutinho, S. 2008. “Self-Efficacy, metacognition, and performance”. *North American Journal of Psychology* 10(1): 165–172.

Hewitt, PM and GS Denny. 2011. “The four-day school week: Impact on student academic performance”. *The Rural Educator* 32(2).

Kalyuga, S, P Chandler, J Tuovinen, and J Sweller. 2001. “When problem solving is superior to studying worked examples.” *Journal of Educational Psychology* 93(3): 579.

Karpicke, JD. 2012. “Retrieval-based learning active retrieval promotes meaningful learning”. *Current Directions in Psychological Science* 21(3): 157–163.

Karpicke, JD and PJ Grimaldi. 2012. “Retrieval-based learning: A perspective for enhancing meaningful learning”. *Educational Psychology Review* 24(3): 401–418.

Karpicke, JD and MA Smith. 2012. “Separate mnemonic effects of retrieval practice and elaborative encoding”. *Journal of Memory and Language* 67(1): 17–29.

Kotovsky, K, JR Hayes, and HA Simon. 1985. “Why are some problems hard? Evidence from Tower of Hanoi”. *Cognitive Psychology* 17(2): 248–294.

Künsch, HR. 1989. “The jackknife and the bootstrap for general stationary observations”. *The Annals of Statistics* 17(3): 1217–1241.

Pope, DG and I Fillmore. 2015. “The impact of time between cognitive tasks on performance: Evidence from advanced placement exams”. *Economics of Education Review 48*: 30–40.

Rafferty, PD. 2013. “Group work in the MBA classroom: Improving pedagogical practice and maximizing positive outcomes with part-time MBA students”. *Journal of Management Education* 37(5): 623–650.

Raman, M, K Mclaughlin, C Violato, A Rostom, JP Allard, and S Coderre. 2010. “Teaching in small portions dispersed over time enhances long-term knowledge retention”. *Medical Teacher* 32(3): 250–255.

Roediger, HL and AC Butler. 2011. “The critical role of retrieval practice in long-term retention”. *Trends in Cognitive Sciences* 15(1): 20–27.

Schneider, W and RM Shiffrin. 1977. “Controlled and automatic human information processing: I. Detection, search, and attention”. *Psychological Review* 84(1): 1.

Seery, MK and R Donnelly. 2012. “The implementation of pre-lecture resources to reduce in-class cognitive load: A case study for higher education chemistry”. *British Journal of Educational Technology* 43(4): 667–677.

Shapiro, SS and MB Wilk. 1965. “An analysis of variance test for normality (complete samples)”. *Biometrika* 52(3/4): 591–611.

Shiffrin, RM and W Schneider. 1977. "Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory". *Psychological Review* 84(2): 127.

Sweller, J. 1988. "Cognitive load during problem solving: Effects on learning". *Cognitive Science* 12(2): 257–285.

Sweller, J. 1994. "Cognitive load theory, learning difficulty, and instructional design". *Learning and Instruction* 4(4): 295–312.

Sweller, J, J Van Merrienboer, and F Paas. 1998. "Cognitive architecture and instructional design". *Educational Psychology Review* 10(3): 251–296.

Van Merrienboer, J and J Sweller. 2005. "Cognitive load theory and complex learning: Recent developments and future directions". *Educational Psychology Review* 17(2): 147–177.

Ward, M and J Sweller. 1990. "Structuring effective worked examples". *Cognition and Instruction* 7(1): 1–39.

Welch, BL. 1947. "The generalization of 'Student's' problem when several different population variances are involved". *Biometrika* 34(1/2): 28–35.

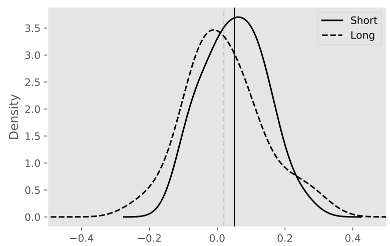
White, H. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". *Econometrica: Journal of the Econometric Society* 48(4): 817–838.

Williams, Rick L. 2000. "A note on robust variance estimation for cluster-correlated data". *Biometrics* 56(2): 645–646.

Yarbrough, R and DA Gilman. 2006. "From five days to four". *Educational Leadership* 64(2): 80–85.

List of Figures

Figure 1: Money and Banking – Exam Score Differences



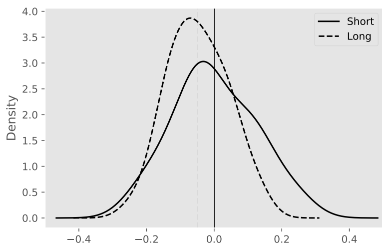
(a) Kernel density estimation (KDE) plot of student exam score differences, the mean value of each class is plotted as a vertical line

	Mean	St. Dev.	25%	75%	Obs.
First Half of Short (S_{F_S})	0.759	0.166	0.636	0.889	87
Second Half of Short (S_{B_S})	0.811	0.151	0.692	0.923	87
First Half of Long (S_{F_L})	0.706	0.203	0.571	0.857	105
Second Half of Long (S_{B_L})	0.726	0.175	0.615	0.846	105
Mean Difference of Differences $((\bar{S}_{B_L} - \bar{S}_{F_L}) - (\bar{S}_{B_S} - \bar{S}_{F_S}))$					−0.031
T-Test of Long vs. Short Differences (P-Value)					0.226

Before performing a KDE or t-test, we first group, then average the differences by the individual student. This has no impact on the mean difference of differences. However, it is a more conservative approach to determining significance given the potential lack of independence across the three observations per student.

(b) Exam scores sorted by when content was taught

Figure 2: Intermediate Macroeconomics – Exam Score Differences



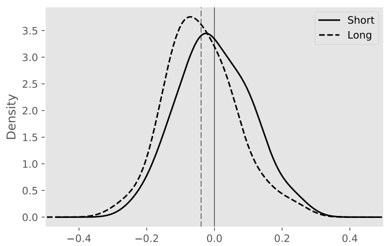
(a) Kernel density estimation (KDE) plot of student exam score differences, the mean value of each class is plotted as a vertical line

	Mean	St. Dev.	25%	75%	Obs.
First Half of Short (S_{F_S})	0.650	0.197	0.500	0.775	87
Second Half of Short (S_{B_S})	0.651	0.163	0.600	0.769	87
First Half of Long (S_{F_L})	0.729	0.166	0.600	0.833	93
Second Half of Long (S_{B_L})	0.681	0.142	0.600	0.769	93
Mean Difference of Difference $((\bar{S}_{B_L} - \bar{S}_{F_L}) - (\bar{S}_{B_S} - \bar{S}_{F_S}))$					-0.049
T-Test of Long vs. Short Differences (P-Value)					0.081

Before performing a KDE or t-test, we group then average the differences by the individual student. This has no impact on the mean difference of differences. However, it is a more conservative approach to determining significance given the potential lack of independence across the three observations per student.

(b) Exam scores by when content was taught

Figure 3: Pooled Exam Score Differences



(a) Kernel density estimation (KDE) plot of adjusted student exam score differences.

	Mean	St. Dev.	25%	75%	Obs.
Adjusted Short Differences	0.000	0.173	−0.116	0.100	174
Adjusted Long Differences	−0.039	0.166	−0.142	0.080	198
Mean Difference of Differences					−0.039
T-Test of Long vs. Short Differences (P-Value)					0.037

For the purposes of this table, we have added approximately 0.051 to all Money and Banking difference observations and approximately 0.001 to all Intermediate Macroeconomics difference observations. This re-centres the distributions such that the short class's mean is at zero while maintaining the original differences. This allows us to pool all class data and compare the mean difference between the short versus long class when the two courses experienced a dissimilarity in the difference in content difficulty.

Before performing a KDE or t-test, we group and then average the differences by the individual student. This has no impact on the mean difference of differences. However, it is a more conservative approach to determining significance given the potential lack of independence across the three observations per student.

(b) Exam scores by when content was taught

List of Tables

Table 1: Money and Banking – Summary Statistics

	Mean	St. Dev.	25%	75%	Obs.
Exam 1 (Short)	0.876	0.084	0.800	0.950	29
Exam 2 (Short)	0.779	0.147	0.700	0.900	29
Exam 3 (Short)	0.712	0.119	0.609	0.783	29
Exam 1 (Long)	0.824	0.117	0.750	0.904	35
Exam 2 (Long)	0.754	0.152	0.673	0.885	35
Exam 3 (Long)	0.636	0.159	0.522	0.761	35

Exam summary statistics for each of the three Money and Banking exams. ‘Short’ is in reference to the daytime section while ‘long’ is in reference to the evening section. Scores are reported as percentage correct [0-1].

Table 2: Money and Banking – OLS Estimation

	Coefficient	Cluster-Robust Standard Errors	P-Value
β (Long Dummy)	-0.031	0.024	0.225
γ_1 (Exam 1 FE)	0.060	0.023	0.010
γ_2 (Exam 2 FE)	0.052	0.024	0.036
γ_3 (Exam 3 FE)	0.042	0.022	0.066
White's LM Test (P-Value)			0.310
Shapiro-Wilk Test (P-Value)			0.589
F-Statistic (P-Value)			0.552
R^2			0.011
Observations			192

The dependent variable is the difference between student i's performance on material taught in the second half of the week and the material taught in the first half of the week ($S_{ieB} - S_{ieF}$). As the long dummy is coded as one for every student in the long (night) class, the coefficient of interest is β.

Table 3: Intermediate Macroeconomics – Summary Statistics

	Mean	St. Dev.	25%	75%	Obs.
Exam 1 (Short)	0.703	0.122	0.600	0.760	29
Exam 2 (Short)	0.670	0.137	0.600	0.760	29
Exam 3 (Short)	0.629	0.157	0.560	0.720	29
Exam 1 (Long)	0.708	0.153	0.600	0.840	31
Exam 2 (Long)	0.751	0.109	0.720	0.840	31
Exam 3 (Long)	0.688	0.109	0.620	0.760	31

Exam summary statistics for each of the three Intermediate Macroeconomics exams. ‘Short’ is in reference to the daytime section while ‘long’ is in reference to the evening section. Scores are reported as percentage correct [0-1].

Table 4: Intermediate Macroeconomics – OLS Estimation

	Coefficient	Cluster-Robust Standard Errors	P-Value
β (Long Dummy)	-0.049	0.027	0.081
γ_1 (Exam 1 FE)	0.058	0.024	0.020
γ_2 (Exam 2 FE)	-0.064	0.025	0.014
γ_3 (Exam 3 FE)	0.008	0.034	0.806
White's LM Test (P-Value)			0.005
Shapiro-Wilk Test (P-Value)			0.222
F-Statistic (P-Value)			0.000
R^2			0.099
Observations			180

The dependent variable is the difference between student i's performance on material taught in the second half of the week and the material taught in the first half of the week ($S_{ieB} - S_{ieF}$). As the long dummy is coded as one for every student in the long (night) class, the coefficient of interest is β.

Table 5: Pooled OLS Estimation

	Coefficient	Cluster-Robust Standard Errors	P-Value
β (Long Dummy)	-0.040	0.019	0.035
γ_{1MB} (Exam 1 MB FE)	0.064	0.020	0.002
γ_{2MB} (Exam 2 MB FE)	0.057	0.024	0.018
γ_{3MB} (Exam 3 MB FE)	0.046	0.022	0.034
γ_{1IM} (Exam 1 IM FE)	0.053	0.021	0.014
γ_{2IM} (Exam 2 IM FE)	-0.069	0.022	0.002
γ_{3IM} (Exam 3 IM FE)	0.004	0.031	0.902
White's LM Test (P-Value)			0.006
Shapiro-Wilk Test (P-Value)			0.460
F-Statistic (P-Value)			0.000
R^2			0.083
Observations			372

The dependent variable is the difference between student i's performance on material taught in the second half of the week and the material taught in the first half of the week ($S_{ieB} - S_{ieF}$). As the long dummy is coded as one for every student in the long (night) class, the coefficient of interest is β.

Table 6: Exam Order Pooled OLS Estimation

	Coefficient	Cluster-Robust Standard Errors	P-Value
β (Long Dummy)	-0.039	0.034	0.263
β_{FE} (Long First Exam Dummy)	0.005	0.045	0.911
β_{SE} (Long Second Exam Dummy)	-0.008	0.040	0.850
γ_{1MB} (Exam 1 MB FE)	0.061	0.023	0.009
γ_{2MB} (Exam 2 MB FE)	0.060	0.026	0.024
γ_{3MB} (Exam 3 MB FE)	0.046	0.026	0.084
γ_{1IM} (Exam 1 IM FE)	0.050	0.023	0.034
γ_{2IM} (Exam 2 IM FE)	-0.065	0.026	0.036
γ_{3IM} (Exam 3 IM FE)	0.003	0.036	0.925
White's LM Test (P-Value)			0.006
Shapiro-Wilk Test (P-Value)			0.462
F-Statistic (P-Value)			0.000
R^2			0.084
Observations			372

The dependent variable is the difference between student i's performance on material taught in the second half of the week and the material taught in the first half of the week ($S_{ieB} - S_{ieF}$). As the long dummy is coded as one for every student in the long (night) class, the coefficient of interest is β. β_{FE} and β_{SE} represent the difference in impact from the long class on the first and second exam from the baseline third exam.

Table 7: Performance Based Pooled OLS Estimation

	Coefficient	Cluster-Robust Standard Errors	P-Value
β (Long Dummy)	-0.037	0.021	0.076
β_{LB} (Bottom×Long)	-0.005	0.037	0.888
β_B (Bottom)	0.042	0.026	0.110
γ_{1MB} (Exam 1 MB FE)	0.046	0.022	0.036
γ_{2MB} (Exam 2 MB FE)	0.037	0.025	0.142
γ_{3MB} (Exam 3 MB FE)	0.027	0.024	0.272
γ_{1IM} (Exam 1 IM FE)	0.034	0.022	0.140
γ_{2IM} (Exam 2 IM FE)	-0.088	0.020	0.000
γ_{3IM} (Exam 3 IM FE)	-0.020	0.030	0.494
White's LM Test (P-Value)			0.023
Shapiro-Wilk Test (P-Value)			0.675
F-Statistic (P-Value)			0.000
R^2			0.096
Observations			372

The dependent variable is the difference between student i's performance on material taught in the second half of the week and the material taught in the first half of the week ($S_{ieB} - S_{ieF}$). As the long dummy is coded as one for every student in the long (night) class, the coefficient of interest is β. β_{LB} captures the difference in impact of the long class on students who scored below the mean on a given exam.