Time Management

Addressing and Assessing Classroom Participation Problems

Cary A. Balser

Abstract While research shows that technology in the classroom has costs, in econometrics (as in other technical courses) computer use is very nearly a necessary condition. Therefore, I used a novel experimental design in concert with SmartSync technology to block cadet use of internet and outlook email on their computers in order to measure the cost of extraneous computer use on: 1) students' self-reported level of attention, 2) students' self-reported level of understanding, 3) performance on low-level, factual learning questions at the end of each lesson and on GRs, 4) performance on higher-level conceptual understanding and application through questions at the end of each lesson and on GRs and through practical exercises, and 5) the accuracy of students' self-reported understanding compared to actual performance. I find that blocking computer distractions at this level has no significant effect on self-reported attention, self-reported understanding, or performance on low-level, factual questions either at the end of a lesson or on GRs. However, I find a .13 effect size of treatment on higher-level questions at the end of lesson and a .058 effect size of treatment effect on the accuracy of self-reported understanding as compared to actual performance. From these results, I suggest that blocking certain computer applications improves short term retention of concepts but that effect on long term retention are unclear and require more research. Finally, this work demonstrates the usefulness of experimental design frameworks in educational research and shows the significance of experiment design in establishing internal validity and testing hypotheses with limited observations. This focus is often not emphasized in educational research which tends to be observational and adds value for researchers considering an educational experiment.

Cary A. Balser, is an assistant professor, Department of Economics and Geosciences, United States Air Force Academy. I would like to especially thank Dr. Lauren Scharff and all the SoTL staff at USAFA, Dr. Nathan Wozny for his support in running the experiment and his invaluable insights, and Dr. Katherine Silz-Carson for her comments and suggestions.

Introduction: As an undergraduate institution with a clear focus on STEM, technology in the classroom is very nearly necessitated by the content in many technical courses. We wish to investigate whether or not introducing software that enables a teacher to restrict access to specific applications and use of computers in class, when paired with immediate assessment surveys, improves student learning and comprehension by enabling technology as an academic tool rather than a distraction.

The use of laptops and other technology in college-level classrooms is a topic of hot debate. Common findings are that students report that having access to laptops in class increased the amount of time spent on non-course material (Zhu et. al., 2011) and that performance is negatively impacted by the off-topic use of computers (Wood et. al., 2012). USAFA is not immune from these concerns. Prior research at USAFA (Fulton, et al. 2011) showed that cadets in three sections of computer science courses (two core level and one programming course) admit to at least occasionally checking emails, checking Facebook or playing games during class lessons. These same cadets also agreed to the comment that "Today's generation is used to multitasking with different technologies and does so well." These self-reports, along with anecdotal evidence from instructors, suggest that multi-tasking on computers during classes at USAFA can be a real challenge.

However, many instructors at the same time believe that the appropriate use of technology is integral to the learning of course material. For example, Wenglinsky (1998) concluded that, "When used properly, computers may serve as important tools for improving student proficiency in mathematics and the overall learning environment of the school." Zhu et. al.(2011) found that students report higher levels of engagement when using technology that is specifically designed to complement the lecture.

How to manage the challenge of multi-tasking is not always clear. Teachers, particularly new instructors, struggle with the choice to allow or disallow technology. On one hand, technology is a part of everyday life for our students and it can be used in ways to make students and teachers more productive. For example, technology can be used to foster discussion or to flip the classroom (Markett et. al., 2006) and in technical classes having technology available allows for students to practice along with or in response to the material in the lecture. On the other hand, technology in the classroom also runs the risk of inappropriate use (i.e. Facebook, Outlook, ESPN, etc.) and can be a distraction and detriment to students. Every teacher has anecdotes of student misuse of technology in the classroom (e.g. I once observed a student looking at cat memes during class), and some teachers find it such a difficult problem that they ban classroom use of laptops all-together. While such a ban may prevent students from being any more distracted than students of the past who did not have technology, it also fails to take advantage of the potentially positive uses of technology in the classroom.

If the goal is to improve teaching and enhance student learning, we must consider why/how multi-tasking is detrimental to learning. Attention is the first step to encoding information to form a memory, i.e. it is the first step in the learning process. Thus, if a student is distracted, then their attention will be drawn away from the material to be learned, even if that information is relatively low-level. For example, Fulton et al. (2011) showed that performance on a basic content multiple-choice and true-false quiz at the end of class was significantly worse for students who checked and responded to email and checked Facebook, with lower performance when students engaged in both email and Facebook. How computer distractions affect higher-level learning (more conceptual / analysis types of learning) compared to basic memorization of terms and content has not been explicitly addressed in the literature that we were able to find. However, we might presume that students will more easily be able to “make up” for distractions in class by memorizing basic content, while it might be more difficult to achieve higher-level learning on their own. Thus, immediate in-class assessments might show equal negative impact of distraction, whereas GR performance might only show negative impact on the higher-level assessment items.

Also related to attention is the research on metacognition that shows students’ awareness of their own learning is also important to achieving mastery of material (Peirce, 2004). Since many students multitask, and multitasking (i.e. using technology during class) affects students’ ability to gather information (Fulton, 2011), students often fail to notice where they have gaps in knowledge (Peirce, 2004). Therefore, another problem of technological (or other distractions) other than the mere learning of information is the students’ ability to accurately understand what they do and do not know in order to tailor study habits and get help as needed. This metacognition of learning is another area in which minimization of distractions may prove helpful to students learning. Therefore, as part of this study, low-stakes (2 pts each) quizzes will be incorporated at the end of class in order to help cadets identify areas of misunderstanding, and to motivate students to attend to class material rather than engaging in off-topic computer use. These in-class assessments will also ask students to rate their level of confidence in their answers, which is a common strategy used to promote metacognitive awareness of learning.

In order to more systematically study the impact of distraction, and to evaluate the effectiveness of a more controlled manner by which to restrict student access to distracting activities on the computer, we propose use of SmartSync software during one section of Econ 365 (Econometrics). Econometrics is an ideal course in which to test effects because computers are necessary to complete the statistical tests and regressions that are taught and to practice and follow along in class. Thus a technology that allows selective blocking of applications while allowing computer use would be ideal. SmartSync is unique because it allows the instructor complete control of any computer in the lab. In particular, SmartSync allows the instructor to selectively shut down any application on any computer or set of computers. This technology also allows immediate communication between the teacher and students, the ready display of any student monitor to the front of the classroom, and the ability to write on or intervene in the screen of any student. This allows a teacher both to call attention to good examples and to correct or note where students are going astray. By using the combination of computer control on some students' computers and interaction on all student computers, instructors can keep students from distracting material, provide real oversight of student use of technology, and monitor students' progress and understanding of the material. The instructor can hopefully leverage all of this information to properly assess students' weaknesses and strengths as well as limit the potential for distraction.

Research Objective and Hypotheses: While research shows that technology in the classroom has costs, in econometrics (as in other technical courses) computer use is very nearly a necessary condition. Therefore, we want to test the impact of 1) low-stakes in-class quizzes to increase awareness of learning and 2) the use of SmartSync to block cadet use of off-topic computer applications. We will examine the impact of these two interventions on:

1. Students' self-reported level of attention
2. Students' self-reported levels of understanding
3. Performance on low-level, factual learning questions at the end of each lesson and on GRs
4. Performance on higher-level conceptual understanding and application through questions at the end of each lesson and on GRs and through practical exercises.
5. The accuracy of students' self-reported understanding compared to actual performance

We hypothesize that students who have limited distractions on their computers (via mechanisms employed by the teacher using SmartSync) will have higher levels of self-reported attention, higher levels of self-reported understanding, better performance on low-level and higher- level questions on the in-class assessments, and have more highly correlated relationships between self-reported understanding and actual performance. Performance on low-level content will not differ on long-term retention assessments (GRs) because students will be able to study and learn that material outside of class even if they did not learn it well during the lesson when they were distracted. However, the performance of those in the treatment group will be better on higher-level GR assessments and practical exercises because students are less able to make up the conceptual understanding lost due to inattentiveness in class.

Research Design: We tested treatments, explained below, using one section of Economics 365 (approximately 20 students) in the national computing lab (4J-17) during each of the 22 lessons during which cadets used computers in class and SmartSync was working. The lab is the only location where SmartSync technology is currently available, and it can no longer be purchased. However, to carry out the course objectives for Econ 365, STATA is needed. Thus we are requesting funds to purchase a lab license for STATA.

For each of the 22 lessons, half of the students will be randomly selected (the treatment group) and blocked from using any application other than a course folder, STATA, and Microsoft Word for note taking, if desired. This randomized assignment means that the treatment group will change from class to class but that, on average, we expect each student to be in the treatment group for about 11 lessons and in the control group for about 11 lessons. This will allow us to hold the fixed effect of a given cadet constant by comparing each cadet's performance under treatment and control and then averaging the differences. However, one might be concerned that cadets may realize that they perform better when in the treatment group and thus adjust their own actions over the course of the semester regardless of assignment to a treatment or control. This would mean that differences towards the end of the semester might appear smaller than the initial effects. We do not have a good way to capture this data, although we might be able to get some feel for this using the questions on attention as the semester goes on, because by controlling for fixed effects of a student we give up the possibility of having a pure control or treatment group. However, I think the fixed effects of each student are likely to be more important than any learning process that might occur. In fact, I'd argue that, in general, the natural tendency is for students to be more distracted or tempted to distraction as the semester goes on. Thus, even if the learning process is occurring, the increased temptation to distraction may minimize that learned effect.

<u>Randomization Technique</u>: Each cadet was be randomly chosen for treatment in half (~11) of the lecture lessons in the course. This was done using excel and a random number generator, in which the 11 largest random numbers for each cadet will denote when a cadet will be in the treatment group. Let us illustrate by taking a cadet with 4 lecture lessons and the following table of randomly generated number for those lessons:

Lesson	1	2	3	4
Cadet X's Random Numbers	0.974782	0.103452	0.525102	0.889218
Treatment?	Y	N	N	Y

Under our randomization technique, Cadet X will be in the treatment group for lessons 1 and 4 (the two largest randomly generated numbers) and in the control group for lessons 2 and 3 (the two smallest randomly generated numbers). This ensured that each cadet is in the treatment group half of the time and that no particular lesson is covered exclusively by the treatment or control group. This randomization is also independent of any choice by a cadet about where to sit or when to show up so that they cannot game the system. However, I found that some issues prevented us from implementing perfectly. First, there were several challenges when trying to get used to the software. Additionally, when the 10th CS completes patches on the system, SmartSync (our treatment technology) often needs to be reset. Overall, the randomization of cadets into treatment and non-treatment has worked well and the cadets have responded well the added challenge of being treated. In order to correct for some of these issues, at the mid semester point I evaluated how many observations of treatment and control I had for each student. I then generated random numbers again for the rest of the semester and assigned treatment and control days in order to balance the number of treatment and control days for each student. This maintained the random selection of which day a student was treated on while trying to ensure that each student had roughly the same number of treatment and control days. In reality, due to absences or other issues, we did have some variation in number of treatment days and days of observation, as seen here in Figures 1 and 2.

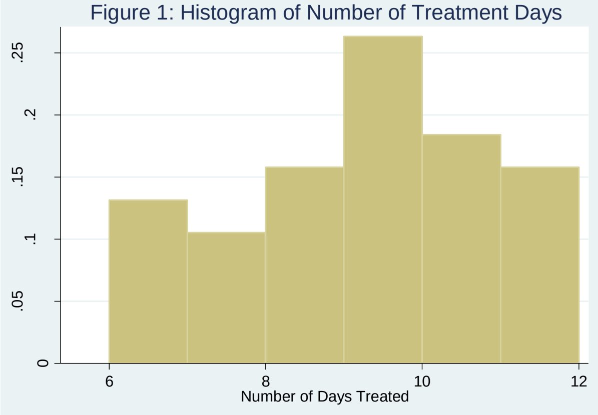


Figure 1: Histogram of Number of Treatment Days

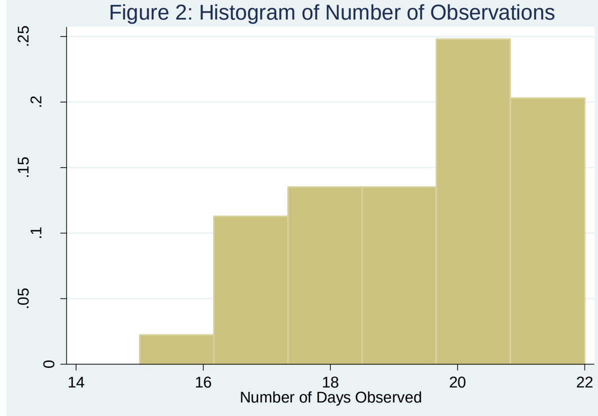


Figure 2: Histogram of Number of Observations

Lesson Layout: The 22 lessons during which the intervention will be carried out each have roughly four parts: posing a policy question/motivation, developing an appropriate econometric model using prior knowledge, adjusting the econometric model to incorporate a new topic/technique or to adjust to a technical concern (i.e. heteroskedasticity), and providing some practice or iteration for the students to complete on their own to reinforce the concepts. Table 1 in Appendix A lays out the general lesson plan components, related instructor actions, appropriate student actions and use of computers, and the benefit that SmartSync can provide during each portion of class to keep students on task with appropriate use of technology. The two types of benefit are the blocking of distractions (e.g. email, Facebook) and the ability for the instructor to monitor student difficulties in order to facilitate student learning (e.g. noting when a student is unable to open a data file).

Assessment Strategy: In order to assess the effect of the SmartSync treatment, all students will receive an end-of-lesson assessment. This assessment will contain questions that are relevant to course material, including three lower-level (multiple choice or true/false questions) and one higher-level (write out or short answer) questions. Students will also be asked to rate confidence in their answers, their perceived level of focus on lesson material and activities, and to identify any technical problems (i.e. "I could not understand the STATA command) they believed might have impacted their learning during that lesson. A sample of an end of lesson assessment for a class concerning dummy variables would be as follows:

EXAMPLE END OF LESSON ASSESSMENT

1. For which of the following would the use of a dummy variable be appropriate?
    a. Grades
    b. Squadron
    c. Race
    d. Age
    e. Nationality
2. Including a dummy variable means that we believe different groups/types will have different ____________:
    a. Intercepts
    b. Slopes
    c. Intercepts and Slopes
    d. Neither Different Intercepts nor Slopes
3. True/False: The constant in our last regression told us the average wage of individuals in the northeast.
4. I believe that being an intercollegiate athlete affects GPA and thus I run the following model: $\mathrm{GPA_i} = \beta_1 + \beta_2 * \mathrm{IC}$. Please interpret the estimate of beta 2.
5. Rate your confidence in your answers from 1 to 10 with 1 being "I am totally unsure of my answers" to 10 being "I'm 100% sure I'm correct."
6. Rate the percent of time (1-100%) that you believed you had focused attention on the lesson material and activities today.
7. Please identify any technical or other problems you had today from the list or choose other and explain:
    a. STATA wouldn't work
    b. I couldn't access the dropbox
    c. I couldn't see other course materials
    d. I could not figure out a command
    e. My connections to the network were down
    f. Other

Questions 1-4 will change based on the lesson material while questions 5-7 will be the same for each lesson. Questions 1-3 in particular will attempt to measure the extent to which lower-level learning was accomplished while question 4 will measure higher level learning. Finally question 5 will be a self-assessment of understanding, question 6 will assess perceived attention to course material, and question 7 will help to parse out and mitigate any other effects that might be causing disruption.

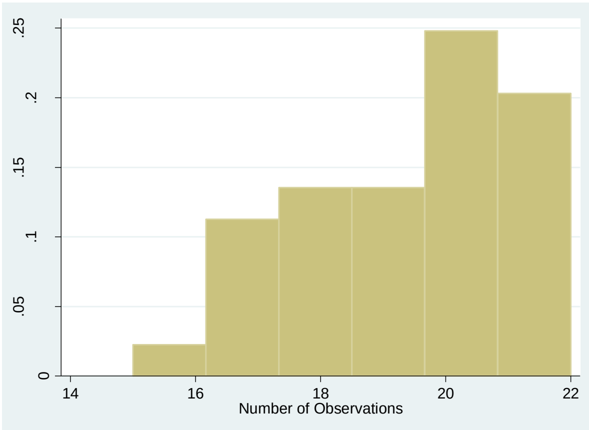
Additionally, we plan to assess long term retention effects by connecting treatment and control groups to performance on specific objectives on tests and quizzes (we will have records for exactly which cadets were under treatment for objectives covered in each class). For example, if cadets X, Y, and Z were under treatment on the day in which dummy variables were covered we can compare them to cadets A, B, and C who were in the control group on the same day by looking at performance on a test question that tests knowledge of dummy variables (i.e. "Suppose we would like to test whether GR scores differ across sections for cadets who have similar scores on analysis 1 and similar scores on analysis 2. Write a null and alternative hypothesis to test this hypothesis, and indicate the model to which your hypotheses apply.). Using this data we can then examine response accuracy between low- and high-level learning questions for both the short (in class) and long term of treatment and control students, and examine how these performance measures correspond to the self-reported confidence and level of focus on the material during the lesson.

Taking the average of all cadets differences will then give us an estimated average difference which we can test either parametrically (using a student's t distribution) or non-parametrically. We can also use descriptive statistics to determine levels of variation for each objective (1-5 as listed above). At prog we will provide an interim report of descriptive statistics regarding effects on overall course performance for those in the treatment and control groups.

Research Findings:

<u>Data</u>: The data I have include the student's treatment, lesson, quiz performance, and self-reported confidence and attention. In total, we have 2,922 observations from the quizzes with a total number of observations between 15 and 22 depending on absences for each student with an overall response rate of 94.7%. Unfortunately, one result is that students who might be most affected, intercollegiate athletes, also happen to miss more class and so I have fewer observations on them. In total I had 9 IC athletes who had an average of 18.55 observations and 27 other students with an average of 19.55 observations. This difference is not significant but IC athletes also had a larger standard deviation of observations. Removing football players, who are ICs but off season in the spring, results in averages of 17.66 and 19.6, respectively, that are significantly different from one another. Figure 3, below, shows a histogram of the

number of observations per student, Figure 4 shows a histogram of response rate per student, and Figure 5 shows a histogram of the response rate by lessons:



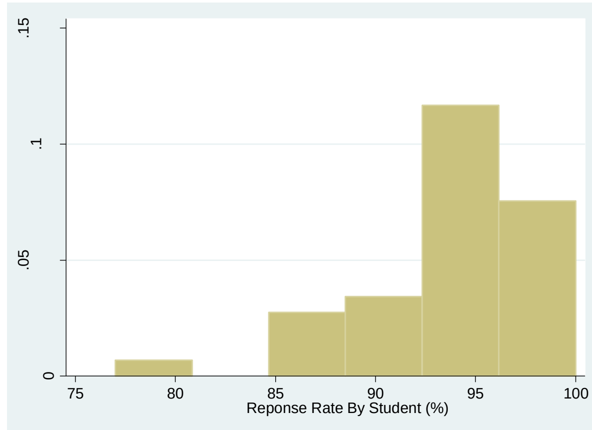


Figure 3: Histogram of Observations

Figure 4: Histogram of Response Rate By Student

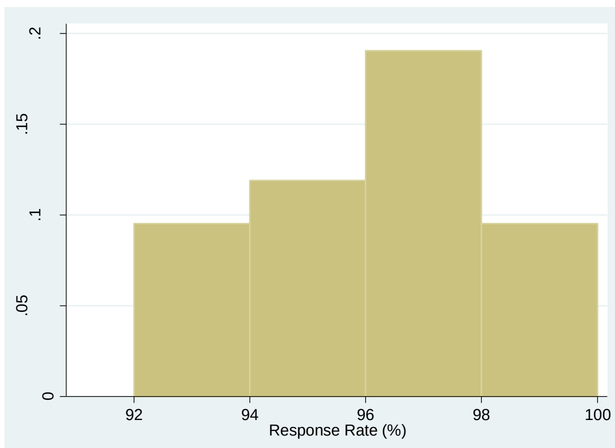


Figure 5: Histogram of Response Rate By Lesson

One quarter of the data points are conceptual and three quarters are low-level. This means that I should have more power in determining any effect on low-level questions, all else equal, if any effect exists. Additionally, referring to figure one, it does not appear that any given student was particularly different from others in terms of their propensity to be treated. This gives a nice balance of treatment and non-treatment for every student in the experiment.

Results

Recall that I predicted that students in a treatment condition will have:

1. Higher self-reported levels of attention
2. Higher self-reported levels of understanding
3. Better performance on low-level, factual learning questions at the end of each lesson

4. No difference in performance on low-level, factual learning questions on GRs
5. Better performance on higher-level conceptual understanding and application questions at the end of each lesson
6. Better performance on higher-level conceptual understanding and application questions on GRs
7. Better accuracy of self-reported understanding as compared to actual performance

I will address each of these hypotheses in order, discussing the results, how they were obtained, and any potential issues regarding the results that I believe exist. +

<u>Self-Reported Attention:</u> Students tend to report high levels of attention but those results tend to vary a lot by which student is reporting. Still, 33% of the time, students report 100% of attention in class is devoted to the material and 72% of responses report attention of 90% or higher. It does not appear, in any of my regressions that treatment has a significant effect on self-reported attention; however, self-reported attention is strongly related to GPA, MPA, and intercollegiate status as seen in Table 1 below. In particular athletes' self-reported attention are, on average, 4.4% points higher than non-intercollegiate athletes with the same GPA, MPA, gender, and within a given lesson. Somewhat surprisingly, those with higher GPA's and MPA's report lower levels of attention. On average, a student with the same treatment type, MPA, gender, IC status during the same lesson who has a 3.5 GPA reports 3.4% points less attention than a student with a 2.5 GPA. Similarly, a student with a 3.5 MPA will report .4% points less attention in class than a student with a 2.5 MPA. One additional consideration was whether treatment might affect students of different types differently. In other words, perhaps students with high GPAs may already pay attention, regardless of blocking and lower GPA students might pay less attention. In this case we'd expect the potential effect of treatment to be much larger based on a student's GPA/MPA but running this specification we find no significant interaction effects of treatment by GPA/MPA. Interestingly, when we leave out student fixed effects our R squared value falls to 0.07, meaning that almost all of our explanatory power is coming from the student fixed effects.

Table 1: Effect of Treatment on Attention in Class

Specification	(1)	(2)	(3)
Treatment	-0.126	-0.088	-11.489
	(0.391)	(0.453)	(8.252)
Grade Point Average	--	-0.034	-0.031
		(0.008)**	(0.011)**
Treatment*GPA	--	--	-0.007
			(0.016)
Military Performance	--	-0.004	-0.006
		(0.001)**	(0.002)**
Treatment*MPA	--	--	0.004
			(0.002)

Intercollegiate Athlete	--	4.426	4.420
		(0.544)**	(0.544)**
Female	--	0.213	0.179
		(0.631)	(0.632)
Fixed Effects Included?			
Lesson	Yes	Yes	Yes
Student	Yes	No	No
Constant	89.498	113.656	118.766
	(65.59)**	(4.406)**	(5.743)**
R^2	0.33	0.07	0.07
N	2,922	2,922	2,922

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

In any case, it does not appear that blocking the internet or email improves self-reported attention levels.

Self-Reported Understanding: I take a similar approach regarding self-reported understanding using the response to the question about how confident students were in their own answers (Note: Students reported confidence on a scale of 1 to 10. I divided those numbers by 10 to estimate the score the student thought they would get. In other words, when a student reported a confidence level of 8, I assumed this meant that they expected an 80%=8/10). In general, students report realistic estimates of expected performance. Like class grades, the average of the distribution is right around 80% (7.8/10 and 1.44 standard deviation) with a long left tail. Here it helps to get a sense of the distribution using Figure 6:

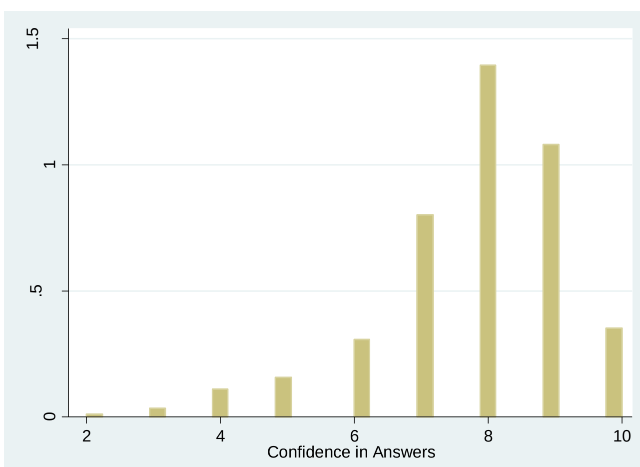


Figure 6: Histogram of Self-Reported Confidence

Here it appears that treatment affects confidence levels as seen in Table 2 below. In particular, the same student in the same lesson reports 8.4% more confidence, on average, in their answers when they are blocked from internet and email access. Further specifications show that females tend to report less confidence in answers (4.7% lower on average) than males with other factors held constant, however, students on IC status are more confident (4.45% higher on average) in their performance, all else held constant. Similar to self-reported attention, student fixed effects still account for a large portion of the variation being explained. Excluding student fixed effects from the regression results in an R squared value of 0.11, a decrease of 0.38. This change emphasizes the importance of the design setup in this research. In other words, although we are explaining relatively little with variables other than fixed effects, we are finding much more precise estimates due to the experimental design that has been implemented.

Table 2: Effect of Treatment on Self-Reported Confidence in Answers (Scores from 0-10)

Specification	(1)	(2)
Treatment Effect	0.084	0.073
	(0.041)*	(0.053)
Grade Point Average	--	-0.001
		(0.001)
Military Performance	--	-0.000
		(0.000)
Intercollegiate Athlete	--	0.445
		(0.063)**
Female	--	-0.474
		(0.073)**
Fixed Effects Included?		
Lesson	Yes	Yes
Student	Yes	No
Constant	8.015	7.843
	(0.143)**	(0.513)**
R^2	0.49	0.11
N	2,922	2,922

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

<u>Performance on Lower-level Questions at the End of Class</u>: First we should note some summary statistics about performance on lower level questions. Students answered 78.9% of lower-level questions correctly on average, under treatment the average was 79.4% which was slightly higher than the 78.5% average when not under treatment. Nearly any specification returns the same result as seen in the below table. Performance on low-level questions has no correlation to treatment and very little of the variation in correct answers is explained, even when including fixed effects. (Note: since our output is either a correct

or incorrect answer, a probit or logit model would also be appropriate but provides very similar results and would be much harder to interpret.)

Table 3: Effects on Low-Level Scores on End of Class Quizzes

Specification	(1)
Treatment Effect	0.003
	(0.019)
Fixed Effects Included?	
Lesson	Yes
Student	Yes
Constant	0.910
	(0.065)**
R^2	0.11
N	2,188

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

Although we expected to see better performance from our treatment group, it is quite possible that such obvious questions are easy to pick up, even when a student is partially distracted. In other words, students may be able to effectively multi-task when the type of material is easy to pick up on even from a cursory listening to the material. Another option, particularly given that the score for both groups is around 75% is that many students are tuned out or distracted independent of the internet/email, and that those tools are just one of many distractions for students. Both seem to have some merit and grounding in reality.

Performance on Lower-level Questions on GRs/Quizzes: In order to make this calculation I matched each GR and Quiz question to a lesson in which that material was covered and whether the question was a lower level question or higher level. I then merged this information with information on students on those days where the questions material was covered to include whether the student was treated on the day that questions material was covered. When holding constant student and lesson fixed effects or consider individual factors, no specification (either that included or did not include fixed effects) found a significant relationship between treatment and performance (see Table 4):

Table 4: Treatment Effect on Average Score for Lower-Level Questions on Tests/Quizzes

Specification	(1)	(2)	(3)	(4)

Treatment	0.031	0.031	0.570	0.663
	(0.020)	(0.020)	(0.377)	(0.380)
GPA		0.001	0.001	
		(0.000)**	(0.000)*	
MPA		-0.000	0.000	
		(0.000)	(0.000)	
Intercollegiate Athlete		-0.004	0.024	
		(0.024)	(0.032)	
Female		0.023	0.005	
		(0.027)	(0.037)	
Interaction Effects				
Treatment and GPA			0.000	0.000
			(0.001)	(0.001)
Treatment and MPA			-0.000	-0.000
			(0.000)	(0.000)*
Treatment and Intercollegiate			-0.060	-0.064
			(0.048)	(0.048)
Treatment and Female			0.045	0.067
			(0.055)	(0.056)
Fixed Effects Included?				
Lesson	Yes	Yes	Yes	Yes
Student	Yes	No	No	Yes
Constant	1.058	0.715	0.479	1.063
	(0.069)**	(0.190)**	(0.254)	(0.069)**
R^2	0.23	0.17	0.18	0.24
N	624	624	624	624

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

Only two significant effects are found (different effect of treatment for those of different MPA's and the effect of GPA). In particular but not surprising is that a 1 point increase in GPA (moving from a 2.5 to 3.5 student) changes average scores by 10% on these non-conceptual or low-level questions holding IC status, gender, MPA and lesson constant. Go figure that those with a GPA, one letter grade higher actually get scores that are one letter grade higher! One final note, even using lesson and student fixed effect, our r-squared values are relatively small, indicating that grades (particularly on lower-level grades) are probably just really noisy measures.

<u>Performance on Higher-level Questions at the End of Class:</u> Relative to lower-level questions, students performed worse on higher-level questions as you would expect. Specifically, students answered such conceptual questions correctly 64.3% of the time, with those under treatment answering correctly 66.9%

of the time and those in the control answering correctly only 62.1% of the time. Running the simplest regression, we estimate that on average, for the same student and lesson, treatment will improve their probability of being right by 6.3% (Note: a probit model predicts marginal effects of 8.2% better under treatment for the average student and lesson, with slightly lower significance—pvalue=.055).

Table 5: Effects on High-Level Scores on End of Class Quizzes

Specification	(1)
Treatment Effect	0.063
	(0.032)*
Fixed Effects Included?	
Lesson	Yes
Student	Yes
Constant	0.516
	(0.111)**
R^2	0.34
N	734

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

Again, no other specifications regarding type or interaction effects have significance. We might also want to know what the estimated effect size is. We can calculate the effect size by dividing the estimated effect above (.063) by the standard deviation of the dependent variable (.479) giving an effect size of .1315.

<u>Performance on Higher-level Questions on GRs/Quizzes:</u> I used the same matching data process described in the lower-level section above to obtain the data. In general, students got 73.0% of the higher-level points on GRs and quizzes (this is much better than scores on similar questions at the end of class). Somewhat surprisingly, treated students scored a 73.8% on average, and those students not treated scored a 72.0% on average. When including student and lesson fixed effects or individual factors, no specification found a significant relationship between treatment and performance (see Table 6):

Table 6: Treatment Effect on Average Score for Higher-Level Questions on Tests/Quizzes

Specification	(1)	(2)	(3)	(4)
Treatment	-0.007	-0.003	0.305	0.296
	(0.019)	(0.019)	(0.359)	(0.350)
GPA		0.002	0.002	

		(0.000)**	(0.000)**	
MPA		-0.000	-0.000	
		(0.000)	(0.000)	
Intercollegiate Athlete		0.045	0.064	
		(0.023)*	(0.030)*	
Female		0.020	-0.003	
		(0.026)	(0.034)	
Interaction Effects				
Treatment and GPA			0.001	0.001
			(0.001)	(0.001)
Treatment and MPA			-0.000	-0.000
			(0.000)	(0.000)
Treatment and Intercollegiate			-0.044	-0.054
			(0.046)	(0.045)
Treatment and Female			0.064	0.060
			(0.053)	(0.052)
Fixed Effects Included?				
Lesson	Yes	Yes	Yes	Yes
Student	Yes	No	No	Yes
Constant	0.757	0.391	0.257	0.753
	(0.059)**	(0.179)*	(0.237)	(0.059)**
R^2	0.51	0.43	0.43	0.51
N	487	487	487	487

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

Although higher-level questions on end of class quizzes saw significant improvements in scores due to treatment, that result does not appear to hold in the long run. Additionally, in the first two regressions without interaction terms, the estimated effects work in the wrong direction as they are negative in two indicating that those treated within lesson and student have lower average scores on high-level questions at GR time. As a sort of sanity check, we do find that those with higher GPA's score better, more specifically, a one point increase in GPA moving from a 2.5 to 3.5 student) changes average scores by 20% on these conceptual or high-level questions holding IC status, gender, MPA and lesson constant. GPA has a differential effect on score based upon the type of question, and, in particular, those with higher GPA's have an even larger advantage on conceptual questions than on non-conceptual ones. We also find that holding fixed effects or other characteristics constant, ICs tend to perform better on conceptual questions than non-IC students.

Accuracy of Self-Reported Confidence to Performance: A primary question might be how self-reported confidence predicts or is correlated to correct answers. For our purposes, I am assuming that self-reported confidence level is equivalent to the student's expected score on that quiz material. What follows from this assumption is that for a 1% increase in self-reported confidence, we expect an equivalent 1% increase in the student's quiz score. We can see in table 7 below the correlations between confidence and performance on low-level and high-level questions under treatment and control. What we find is that the relationship between confidence and treatment is stronger under treatment for both high and low level questions.

Table 7: Relationship between Self-Reported Confidence and % Correct

Specification	(1)	(2)	(3)	(4)
Low-Level	Yes	Yes	No	No
Treated	No	Yes	No	Yes
Confidence in Answers	0.014	0.027	0.037	0.054
	(0.012)	(0.014)	(0.020)	(0.024)*
Fixed Effects Included?				
Lesson	Yes	Yes	Yes	Yes
Student	Yes	Yes	Yes	Yes
Constant	0.956	0.509	0.126	0.211
	(0.126)**	(0.154)**	(0.213)	(0.266)
R^2	0.12	0.15	0.40	0.38
N	1,189	999	399	335

Effect size above with standard errors below in parentheses; * for 5% significance; ** for 1% significance

Additionally, the only significant result is when there are high-level questions for students under treatment. In that case, a one point increase (10% increase) in confidence increases the % correct by .5%. This is a relatively small change even though it is significant which shows that there appears to be a generally correct tendency for students to report higher confidence when they are actually performing better. However, this relationship does not tell us whether students reported confidence levels accurately predict their performance.

In order to investigate how accurate students are in their own confidence I created two new measures, one for the short term "dissonance" students seem to have and one for long term "dissonance". For the short term, I took the score of a student on their end of class and subtracted that score from the student's confidence level on that day for each student and lesson pair. For example:

$$
\begin{aligned}
&Short\ Run\ Dissonance\ (for\ Student\ X\ in\ Lesson\ X) \\
&\qquad = Lesson\ X\ Confidence\ in\ Answers \\
&\qquad - Overall\ Score\ on\ End\ of\ Class\ Quiz\ for\ Lesson\ X
\end{aligned}
$$

This gives a difference between actual short run and expected performance (as measured by self-reported confidence) where positive values indicate a student is overestimating their knowledge and negative values indicate a student is underestimating their knowledge. In order to do this for the long run, however, I had to take a weighted score of conceptual and non-conceptual questions. Because my confidence levels related to how confident students were on four questions (three of which were non-conceptual and one of which was conceptual) and because I'm assuming they gave equal weight to each question in reporting confidence, I used the following formula for a given Lesson and Student pair from related GR/quiz scores to get a long term score for that lesson:

$$
\begin{aligned}
&Long\ Run\ Score \\
&\qquad = Related\ Non\ Conceptual\ GR\ or\ Quiz\ \%\ Correct * .75 \\
&\qquad + Related\ Conceptual\ GR\ or\ Quiz\ \%\ Correct * .25
\end{aligned}
$$

For a given student and lesson, I then took confidence levels and subtracted them from the score of that student on the end of class quiz. For example:

$$Long\ Run\ Dissonance = Lesson\ X\ Confidence\ in\ Answers - Long\ Run\ Score\ for\ Lesson\ X$$

Again, this gives a difference between actual long run and expected performance (as measured by self-reported confidence) where positive values indicate a student is overestimating their knowledge and negative values indicate a student is underestimating their knowledge. However, it is worth noting that using this process meant I had to drop Lessons when I did not have both a conceptual or non-conceptual question. I also had to drop Confidence Scores for lessons where I did not have a related set of GR questions at all. This trimmed the data set down significantly, from 734 short run observations to 452 long run observations. Nevertheless the summary statistics show that students overestimate their knowledge in both the short and long run, although they tend to overestimate more in the long run. This is confirmed by the following set of histograms and density plots. Is a histogram of dissonance scores in the long run and on the right a histogram of dissonance scores in the short run:

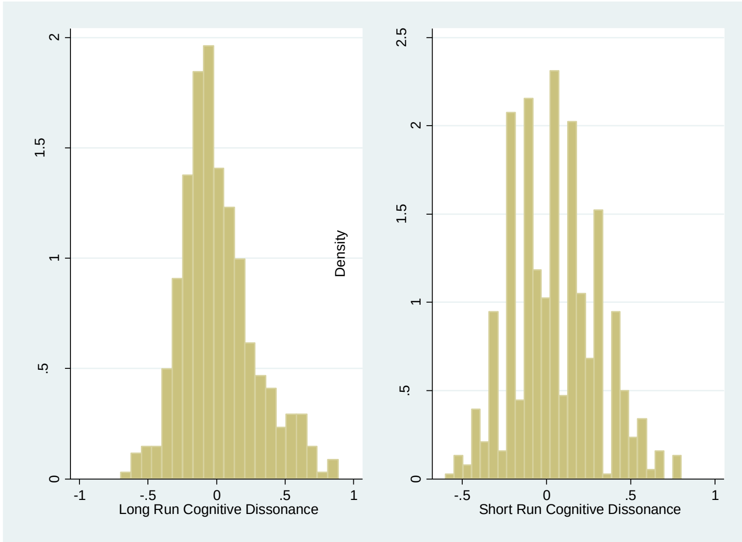
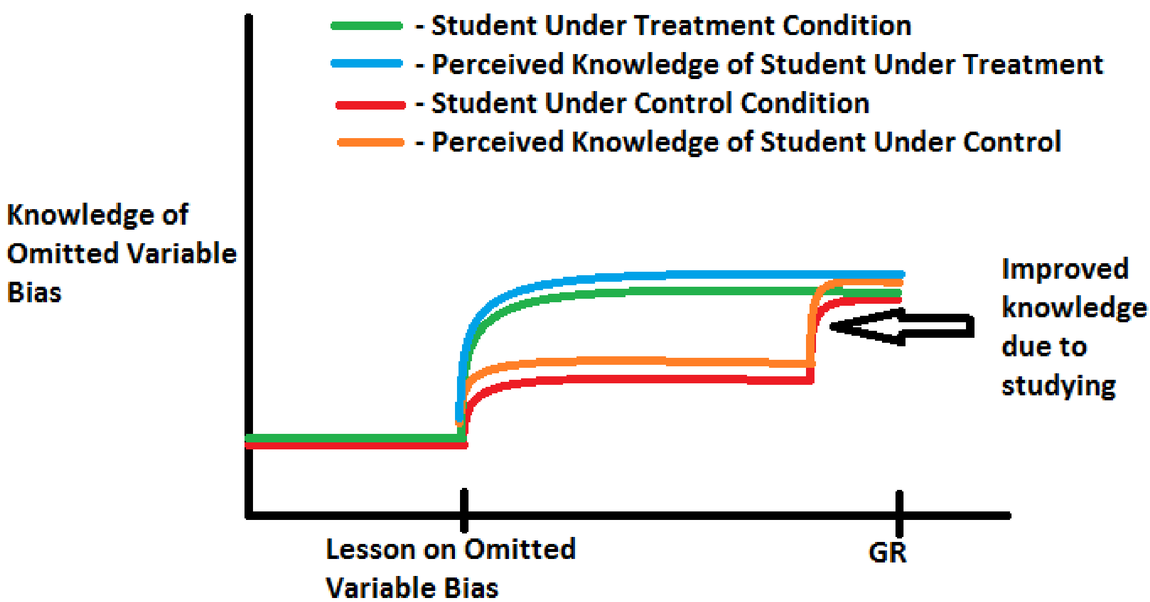


Figure 7: Cognitive Dissonance Scores in the Long Run and Short Run

What is most striking here is just how close the center of the distributions are to zero, although both tend to have long right tails. Summary statistics show that, on average, students overestimate their knowledge by about 1.2% in the long run and by about 5.3% in the short run although both have approximately the same standard deviation. Additionally, while regressions show no significant effects of treatment on cognitive dissonance in either the short or long run, I find cognitive dissonance decreases by 7% points and 14.6% points for females relative to males in the short and long run respectively. Similarly, for each one point in GPA for each additional point of GPA, a student will have cognitive dissonance scores that are 5.6% points lower and 7.3% points lower in the short and long run respectively. Thus, the higher a student's GPA, the more they underestimate their actual performance, on average.

Interpretation of Findings and Conclusion: The bottom line is that blocking distractions via computers does improve learning in the short run for more complicated concepts. However, those gains do not appear to hold in the long run based on GR scores, appear to have no effect on more rote learning, and appear to have no effect on cognitive dissonance as regards their knowledge of material. The most interesting follow up question is why don't short run gains due to treatment translate to long term gains? One possible explanation is just that GR scores are not related to in class learning. This is a pessimistic view that would assert that classroom time does not matter for test performance or learning outcomes. It also is not in line with our knowledge of the effects of classroom instruction. Another possibility is that the related GR questions do not actually line up with the lesson material from the days I am proposing they line up with. There is probably some truth to this because GR questions tend to cover more than one concept and more than one topic, making it hard to isolate the effect on long term retention due to one lesson or to the treatment condition during a given lesson. Another plausible explanation is that students who were under control knew they were not paying attention as much, or at least that they had a worse understanding of the knowledge, and accordingly allocated more time to studying the topics for which they were not blocked. For example, in the picture below we see four lines representing the knowledge and perceived (self-assessed) knowledge of two students on the topic of omitted variable bias. In this



story the student under treatment, call him Bob, learns much more than the student under control, call him Steve, during the course of the lesson (we assume Bob and Steve are the same in every other respect). This is consistent with our short term findings of the effect of treatment. You'll notice that both Bob and Steve perceive themselves as understanding slightly more than they actually do on the topic but that the differences between perceived and actual knowledge is the same for Bob and Steve. In other words, both are equally aware of their true understanding regardless of treatment which is consistent with my findings. Now as the GR approaches, Bob is comfortable with his level of knowledge on the topic while Steve realizes he needs to "catch up." As a result, Steve devotes study time to the concepts of omitted variable bias and "catches up" resulting in roughly the same GR score as Bob (consistent with our findings of no differences between treatment and control groups on the GR). So, what is the take away from this particular version of the story? It may be that by treating students, we make class time more effective and limit the amount of time they need to study. Which leaves me to ask which of these stories is accurate? To what extent do any of the stories above explain our results? To test the effect of treatment on long term retention, we need a more focused experiment to tease out these effects. Perhaps even survey data would be appropriate here by asking students how much time they spent studying each objective and relating that to their treatment and control days. If the story about Bob and Steve is even partially accurate, then there is a huge advantage to students by being blocked from internet in class because it will limit the amount of time they need to spend studying. One final note is that the random assignment may actually have some effect on the result we saw. If students always expected to be blocked from the internet, it is possible they would change their behaviors in ways they did not when they don't know if they will be blocked or not. Thus a heavy handed approach to blocking the internet may not achieve the same results we see here. The best strategy may just be to randomly (or strategically) implement blocking by student, lesson, or both.

Supporting Documents and References:

Bugeja, Michael. "Distractions in the Wireless Classroom." *The Chronicle of Higher Education*. N.p., 26 Jan. 2007. Web. 03 Dec. 2014.

Bugeja, Michael J. "Facing the Facebook." *The Chronicle of Higher Education*. N.p., 23 Jan. 2006. Web. 03 Dec. 2014.

Fang, Berlin. "From Distraction to Engagement: Wireless Devices in the Classroom (EDUCAUSE Quarterly) | EDUCAUSE.edu." *From Distraction to Engagement: Wireless Devices in the Classroom (EDUCAUSE Quarterly) | EDUCAUSE.edu*. EDUCAUSE, 22 Dec. 2009. Web. 02 Dec. 2014.

Fisher, Beth, and Gina Frey. "Research on In-Class Use of Laptops and Other Devices: Effects on Students' Learning and Attention." *Teaching With Technology*. The Teaching Center Journal, 24 Sept. 2014. Web.

Fulton, Steven, Dino Schweitzer, Lauren Scharff, and Jeff Boleng. *Demonstrating the Impact of Multitasking in the Classroom*. Proc. of 41st ASEE/IEEE Frontiers in Education Conference, Rapid City, SD. N.p.: n.p., 2011. F2J-1-2J-6. Print.

Markett, C., I. Arnedillo Sánchez, S. Weber, and B. Tangney. "Using Short Message Service to Encourage Interactivity in the Classroom." *Computers & Education* 46.3 (2006): 280-93. Web.

Peirce, William. *Academic.pgcc.edu*. Proc. of METACOGNITION: Study Strategies, Monitoring, and Motivation. Prince George's Community College, 17 Nov. 2004. Web.

Wenglinsky, Harold. "Does It Compute? The Relationship between Educational Technology and Student Achievement in Mathematics." *Does It Compute? The Relationship between*

Educational Technology and Student Achievement in Mathematics. Educational Testing Service, 1998. Web.

Wood, Eileen, Lucia Zivcakova, Petrice Gentile, Karin Archer, Domenica De Pasquale, and Amanda Nosko. "Examining the Impact of Off-task Multi-tasking with Technology on Real-time Classroom Learning." *Computers & Education* 58.1 (2012): 365-74. *Elsevier*. Web.

Zhu, Erping, Matthew Kaplan, R. C. Dershimer, and Inger Bergom. "USE OF LAPTOPS IN THE CLASSROOM: RESEARCH AND BEST PRACTICES." USE OF LAPTOPS IN THE CLASSROOM: RESEARCH AND BEST PRACTICES. University of Michigan: Center For Research on Learning and Teaching, 2011. Web.