

Evaluating Student Performance on Computer-Based versus Hand-Written Exams: Evidence from a Field Experiment in the Classroom^{*†}

Besir Ceka

L. Richardson King Assistant Professor
Davidson College

Andrew J. O'Geen, Associate Professor
Davidson College

Pre-print version. Forthcoming in *PS: Political Science & Politics*.

Abstract

The use of course management software such as Blackboard, Moodle, and Canvas has become ubiquitous at all levels of education in the United States. One potentially useful feature of these products is the ability for instructors to administer assessments like quizzes and tests, which are flexible, easy to customize, quick and efficient to grade. While computer-based assessments offer clear advantages, instructors might worry about their effect on student performance. This paper evaluates whether student-based performance differs between hand-written and computer-based exams through a randomized field experiment conducted in a research methods course. Overall our findings suggest a significant improvement in student performance on computer-based exams that is driven primarily by the relative ease of producing thorough responses on the computer versus by hand.

* Besir Ceka is the L. Richardson King Assistant Professor in the political science department at Davidson College. He can be reached at [beceka@davidson.edu](mailto:beceka@ davidson.edu). Andrew J. O'Geen is an associate professor in the political science department at Davidson College. He can be reached at anogeen@davidson.edu.

† We would like to thank Shelley Rigger and Silvana Toska for the initial fruitful discussion where the idea for this paper originated, and Jess Good and Laura Sockol for helpful suggestions on research design. We would also like to thank members of the Experimental Methods Group at the European University Institute in Florence, Italy, and especially Klarita Gërxhani, Eva Zschirnt, and Marc Helbling for their helpful suggestions. We are also grateful to Bilyana Petrova and Melody Crowder-Meyer for their incisive comments on an earlier draft. Finally, we would like to thank Victoria Frost, Daphne Asteriadis, and Cole Riggan for their research assistance, the students that participated in the experiment, and three anonymous reviewers for their helpful feedback.

Computer-based assessment offers many obvious advantages over hand-written assessment. Independent of the method of administration, typed answers are much easier for instructors to read and thus grading and evaluation can be done more efficiently. In addition, typing can allow students to produce more in response to questions and can ease the editing process, making answers potentially more complete and more accurate. Computerizing assessment allows instructors to administer quizzes and tests which are flexible, easy to customize, and quick and efficient to grade.

While computer-based assessments offer clear advantages, instructors might worry about their effect on student performance. The literature on the equivalence of computer- and paper-based assessments has yielded mixed results. Some studies find that computer test-taking enhances student performance, others find that paper is better, while others still find equivalence between the two test modes (see Noyes and Garland 2008 for an overview). The contradictory findings are in part due to the variety of assessments used (e.g. multiple-choice questions, essays etc.), the different thresholds for equivalence, and the different research designs employed (see Albers, Boevé, and Meijer 2015). This implies that the topic is a moving target warranting further study.

The goals of this paper are twofold. First, we evaluate whether student-based performance differs between hand-written and computer-based exams through a randomized field experiment conducted in our research methods course. Finding a significant difference, we then investigate potential causes of that difference. Overall our findings suggest nearly a full grade increase in student performance on computer-based exams that is driven primarily by the relative ease of producing thorough responses on the computer versus by hand.

Our study contributes to the existing literature on test-taking by employing a particularly suitable but relatively underutilized experimental design in a naturalistic setting. Moreover, our

investigation goes beyond simply evaluating the equivalence of test modes and provides an explanation for why performance might be higher in computer-based exams. Finally, by using student-generated data to answer a question of direct interest to students and faculty alike we provide another tool for engaging students with research methods in political science (Adriaensen, Kerremans, and Sloomaeckers 2015; Rosen 2018).

Context

We conducted our IRB-approved experimental design in two sections of a required introductory course in methods and statistics in Political Science in the fall of 2017 at a small liberal arts institution. The sections, capped at 30 and yielding a total enrollment of 59 students, met three times per week one after the other in the same classroom.¹ Each instructor taught a single section of 29 or 30 students. The course sought to introduce students to the fundamentals of social science research, including statistical analysis and computer programming. The two instructors used the same syllabus for both sections and coordinated each class meeting to cover the same material.

Experimental Design

We used two equally weighted 50-minute in-class exams to evaluate whether the test-taking mode affects student performance.² Each of these exams tested students on the quantitative part of the course and required them to produce several short-answer/essay responses. Although the nature of the course material is cumulative, the two exams were designed to test for different material. The exam questions and instructions were identical across sections and were administered in the computer lab where the class met during the semester. The first two rows of the classroom were reserved for students that took their exams on paper, while the last two rows were reserved for those who took them on a lab computer. Thus, we essentially ran two parallel

experiments, one in each section, using two separate exams and two treatments—(1) the computer-based version of each exam administered via learning management software (Moodle) and (2) the paper-based version of each exam taken with pen and paper.

In each section, students were randomly assigned to one of the two sequences of test-taking—paper first or computer first—and then switched to the other test-taking mode in the second exam. They were informed about their test-taking sequence before they took the exams. For example, if a student was randomly assigned to take the first exam using a computer, she took the second exam using paper and vice versa. More technically, we used a counterbalanced within-subjects design, also known as a 2x2 crossover design, with two periods (exam 1 and 2) and two treatments (computer and paper) (see Figure 1).

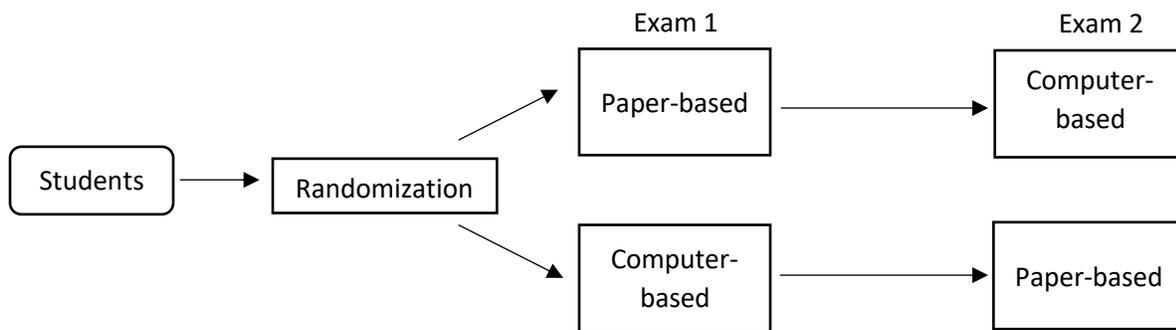


Figure 1. Schematic representations of the within-subject experimental design

This experimental design is particularly suitable for our research question (Albers, Boevé, and Meijer 2015). First, it treats each subject as one's own control, which effectively removes all the differences between subjects in calculating average treatment effects (Jones and Kenward 2014). Thus, the treatment effects are within-subject and account for all the between-student differences that could have arisen due to individual factors and due to instructor differences.

Second, our counterbalanced design is uniform within periods and sequences, which means that each treatment appears an equal number of times within each period and each sequence. This allows us to remove period and sequence effects that could be confounding factors. For example, it could be that students become more comfortable with the material over the course of the semester and perform better in later exams than in earlier ones. Or it could be the case that the test-taking sequence, computer or paper first, affects performance. By randomly assigning students to each test-taking mode and then switching them to the other mode, we account for both of these effects.

Finally, crossover experiments require fewer subjects, which makes them statistically more efficient at estimating treatment effects. In fact, one needs a far smaller sample size, often less than half of that of a traditional parallel design, to achieve the same statistical power (Li et al. 2015). Given the relatively small number of students in our classes, this is a crucial advantage.³

Despite these advantages of our experimental design, we were acutely aware of the potential for bias to enter our analysis through differences in grading both within and between instructors. To avoid within-instructor grading bias that could stem from preference for one test mode versus another, we employed three research assistants who transcribed the hand-written answers and put them in a spreadsheet. We then added the answers from the computer exams to the same spreadsheet and graded answers in our own sections anonymously. To avoid between instructor bias, we employed two strategies. First, the crossover design effectively removes differences in performance between students that could arise due to the teaching or grading style of the instructor because treatment effects are estimated for each individual student across treatments. Second, we discussed our grading criteria in-depth and agreed on suitable answers for each question.⁴ The formal way to test for instructor effects is to introduce an indicator variable for section in the analysis of variance and see whether there is an interaction between sections and

the direct treatment effect (Jones and Kenward 2014, 62–63). Our analysis shows that there is no evidence for such an interaction, further warranting the combination of data from both sections into one analysis.

Another disadvantage of crossover designs is the potential for the treatment at time t to affect the outcome at time $t+1$. In other words, if a subject receives treatment A in the first period and treatment B in the second period, the outcome measured at the second period could be in part due to the lingering effect of treatment A. This carryover effect is of particular concern in pharmaceutical trials where the administration of a certain drug in the first period could affect the outcome of interest at a later period. To reduce such carryover effects, crossover designs often include washout periods.⁵ A carryover effect, however, is unlikely to occur in our application because there is no reason to expect that using one test-taking method in the first exam should have any lingering effect on student performance later on (see Senn 2002 on carryover effects). However, given our main findings, one could argue that the students who took the first exam on a computer might feel more confident in their abilities going forward, and hence perform better on the second exam conducted on paper. That clearly did not happen, further reducing concerns about carryover effects.

Finally, our experiment meets all the criteria for field experiments. Field experiments are conducted in naturalistic settings, which gives them higher external validity than lab experiments.⁶ In our case, the treatments resemble real test-taking modes, the participants are students who generally take tests, the relevant context is the classroom, and the outcome measures are actual test scores (Gerber and Green 2012, 10–11). The high-stake nature of the exams and the setting in which they are taken thus closely mirror the test-taking experience of the average student. In

addition, our within-subject crossover design greatly reduces concerns about confounders and internal validity, which often plague field experiments.

Results

To properly estimate the treatment effect in crossover designs, we need to account for the fact that observations are paired. In other words, since student A's performance in period 1 is obviously related to her performance in period 2, we need a statistical test that accounts for the repeated measure nature of the data. To do this, we use analysis of variance (ANOVA) for a 2x2 crossover study.

Table 1. Results for the two-treatment, two-period crossover design

<i>Treatment sequence</i>	Exam		Within-student difference: Computer - Paper
	1	2	
<i>Paper then Computer</i>			
<i>Mean (sd)</i>	85.37 (7.93)	79.60 (15.38)	-5.78 (13)
<i>Sample size</i>	28	28	28
<i>Computer then Paper</i>			
<i>Mean (sd)</i>	83.9 (9.04)	65.41 (16.64)	18.49 (14.33)
<i>Sample size</i>	25	25	25
<i>Treatment effect</i>			
<i>Mean (sd) and 95 CI</i>	-	-	6.36 (1.87) [2.60 to 10.12]
<i>Sample size</i>	-	-	53
<i>T-test</i>	-	-	P-value: 0.001

Table 1 reports summary statistics of the within-student difference with respect to exam scores across sequences of test-taking as well as the total treatment effect from ANOVA controlling for period effects. The second exam proved more difficult than the first, with average scores being about 12 points lower, providing us with evidence of a significant period effect. However, average scores fell far less for the students who took their second exam on computer than for those who took it on paper. Figure 2, which plots student scores for each period by sequence of treatment, confirms this impression. The solid lines indicate each individual student

in the sequence who managed to improve their score between exams. While students tended to perform better in the first exam than in the second, it is clear that more students improved their score when they took the second exam on a computer rather than on paper. Moreover, there was a steeper drop in performance for students who took the second exam on paper than on computers, further implying a treatment effect.

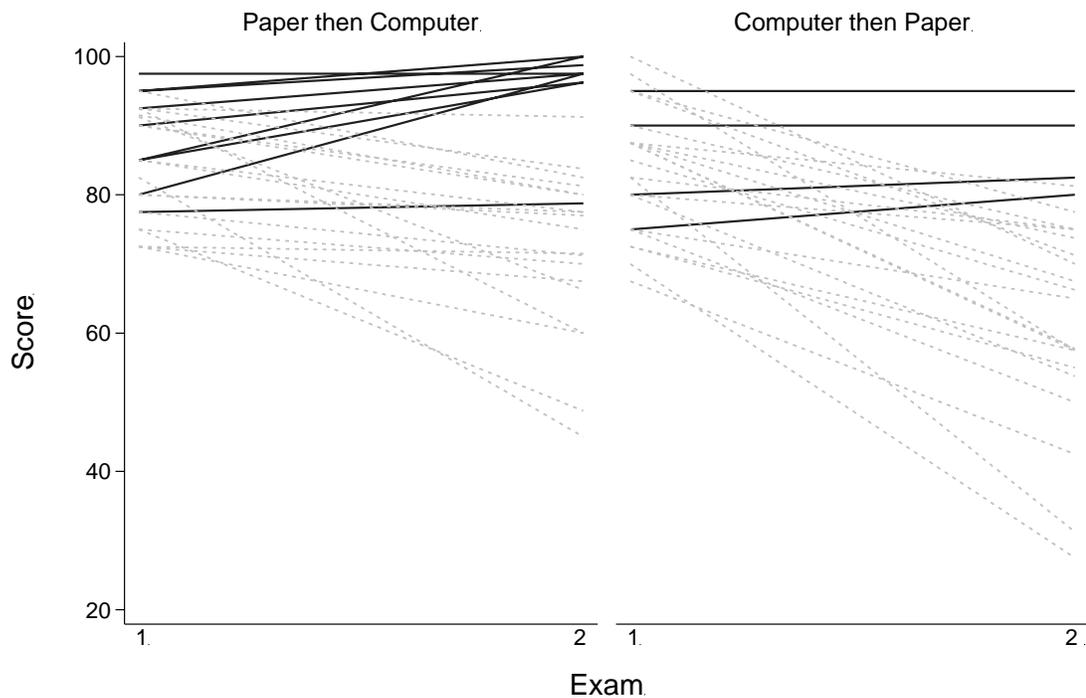


Figure 2. Profile plots for student scores for each period by sequence of treatment

The within-subject treatment effect and the corresponding 95% confidence interval for taking the exams on computer versus paper are shown at the bottom of Table 1. Accounting for the period effect, students scored about 6 points higher when they took their exam on computer than on paper. This effect is statistically significant at the 0.001 level. This means that computer test-taking led to nearly a full letter grade increase in student performance, which is a substantively

large effect. In short, there is fairly strong evidence that taking exams on computers as opposed to paper not only does not hinder student performance but actually enhances it.

Further discussion and analysis

Why is there a positive treatment effect for computer-based tests? These results were surprising to us, so we decided to probe them further by asking our students to fill out a quick online survey on note-taking preferences at the end of the course (see appendix A2). The idea was to see whether a preference for note-taking on a computer helps with the retrieval of information and, thus performance, during exams also taken on a computer. Of the students who filled out the survey, almost four-fifths (43) preferred paper and only about one-fifth (12) preferred computers for note-taking. The ANOVA analysis with interaction effects between treatment and note-taking preference for computers showed no statistically significant results. In other words, the boost in performance from computer test-taking does not depend on a preference for note-taking on computers.

Another alternative explanation is that taking tests on a computer allows one to provide more comprehensive and higher quality answers due to the higher speed of typing and ease of editing (see Goldberg, Russell, and Cook 2003). Indeed, students taking the exams on paper wrote on average 488 words while those taking them on a computer wrote 655 words (see Table 2 for additional information).

Table 2. Descriptive statistics for the total number of words by treatment

Total number of words	Mean	Sd	Median	Min.	Max.
Paper	488	134	472	261	861
Computer	655	184	647	285	1081
Overall	571	181	537	261	1081

To what extent is the observed treatment effect due to the fact that students can write better and more comprehensive answers with computers? To explore this possibility, Figure 3 plots students' scores over the range of the number of words they wrote. The figure shows a generally positive relationship between the total number of words written per exam and the total score, but this relationship does not seem to be linear. Indeed, a quadratic line seems to fit the data better.

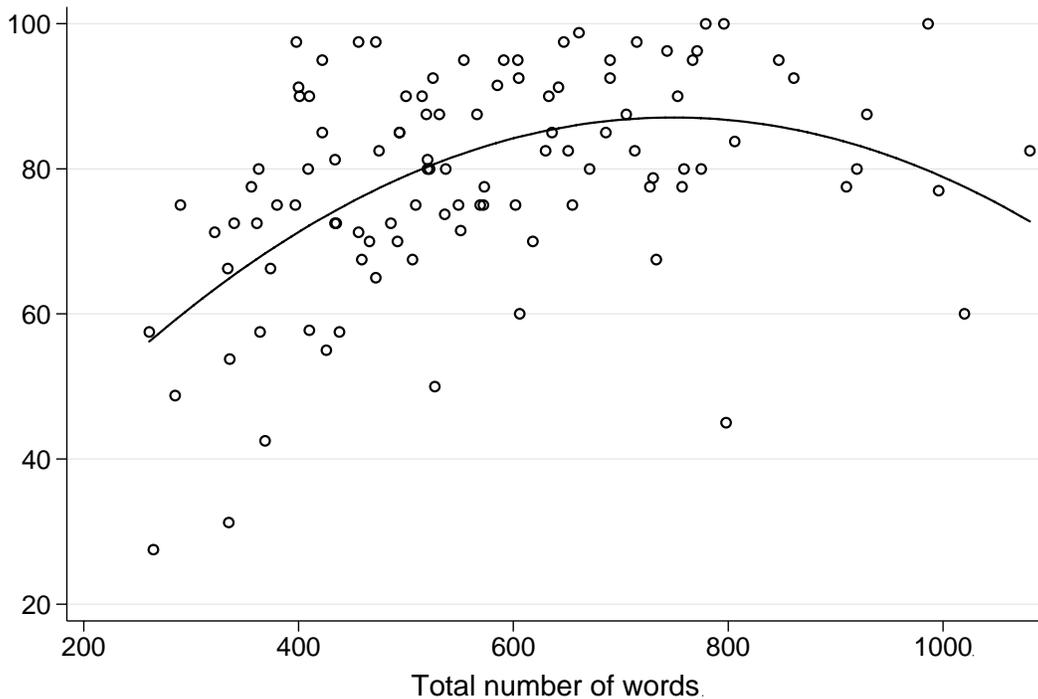


Figure 3. Scatterplot with a quadratic fitted line for student scores against the total number of words written

We were unsurprised by this result, given that writing longer answers can be a double-edged sword. While more comprehensive answers can give students a chance to display the depth of their knowledge, they can also make gaps in understanding more transparent.

To further investigate, we conducted additional ANOVA analyses by including the total number of words and its square as covariates in the base ANOVA model. The regression coefficient for the total number of words variable is positive and statistically significant at the 0.05

level, while its square is negative and statistically significant as well, confirming the inverted U-shape of the quadratic fitted line in Figure 3. Most interestingly, controlling for the effect of the total number of words renders the positive treatment effect for computers statistically insignificant. This suggests that test-taking on computers increases student performance mostly by allowing them to write longer and more comprehensive answers. However, it is quite likely that the typed answers were also better edited and more polished because they lend themselves more easily to proofreading, and correcting for spelling and grammar errors, than do handwritten exams. While we cannot empirically test for this effect with our data, its presence would further support the use of computer-based assessment.

Conclusion

Our main motivation for this study was to investigate whether student performance differs between hand-written and computer-based exams. This is a practical concern that we and other colleagues share about the role of technology in the classroom in general, and the effect of computer test-taking on performance in particular.

Our findings suggest that students tended to perform better when they took their exams on a computer rather than on paper. This effect was due to the more comprehensive answers that they provided when given the opportunity to type rather than write. Computer-based assessment offers benefits to instructors, too. They allow for quicker grading because (1) it is easier to read typed answer than handwritten ones and because (2) it takes less mental effort to grade by question than by exam, which learning management software like Moodle often make very simple to do. We estimate that it takes us half the time to grade the same answers on a computer than on paper. Moreover, in-class computer-based exams can be easily administered in any classroom as long as students have access to laptops and can bring them to class. Coupled with the elimination of the

halo effect that accompanies grading student exams anonymously by question (Kahneman 2013, 83-84), we believe that computer-based tests offer clear advantages to both students and instructors alike.

Our study undoubtedly has a number of potential limitations. One is the relatively small number of students involved in the study. Although the crossover design alleviated this, a larger sample size would provide enough statistical power to detect even smaller effects. Second, the exams were administered as part of a required methods course, so future research could investigate whether the results hold for other types of courses in different subjects and disciplines, and using other course management platforms. Finally, the exams were timed and relied on short answer/essay questions, so it remains to be seen whether our findings apply to other types of questions and time limits.

A secondary motivation for conducting our study was that we wanted to design a built-in field experiment in a methods course that served to make the various steps of research design and analysis more tangible to students. Our hope was that working with data they helped generate would increase student interest in and engagement with a material that oftentimes is outside of “students’ sphere of interest” (Adriaensen, Kerremans, and Slootmaeckers 2015, 4) and causes unusually high student anxiety (Bernstein and Allen 2013).

A variety of approaches have been proposed to pique student interest including flipping classrooms (Touchton 2015), assigning relevant practical examples from existing research, and even using examples from popular culture (Burkley and Burkley 2009). We provide an additional solution to this problem by using student-generated data to address a substantive issue of direct interest to students (see also Rosen 2018). In our example, the subject of academic performance and the setting of a required methods course offered an ideal opportunity to engage students with

a research question that was of immediate relevance to them. The experiment served as a built-in example that we were able to thread throughout all aspects of the semester-long course: from developing research questions, to operationalizing concepts, to designing experiments, to data analysis and reporting (see appendix for more on this).

In sum, evaluating whether test-taking modes affect student performance is highly valuable in an age when technology is changing the way students and instructors interact in almost every aspect of higher education. Doing this evaluation in a systematic manner within the confines of a research methods course provides an excellent pedagogical tool to engage students with the material and to bring it closer to their “sphere of interest.”

The question we chose to investigate and the method we employed represent just one possibility. Others could implement the same principle—i.e. enhancing student engagement with research methods through the collection and analysis of student-generated data—by using different questions of interests and other methods commonly used in political science.

References

- Adriaensen, Johan, Bart Kerremans, and Koen Sloomaeckers. 2015. "Editors' Introduction to the Thematic Issue: Mad about Methods? Teaching Research Methods in Political Science." *Journal of Political Science Education* 11 (1): 1–10.
- Albers, Casper J., Anja J. Boevé, and Rob R. Meijer. 2015. "A Critique to Akdemir and Oguz (2008): Methodological and Statistical Issues to Consider When Conducting Educational Experiments." *Computers & Education* 87 (September): 238–42.
- Bernstein, Jeffrey L., and Brooke Thomas Allen. 2013. "Overcoming Methods Anxiety: Qualitative First, Quantitative Next, Frequent Feedback Along the Way." *Journal of Political Science Education* 9 (1): 1–15.
- Burkley, Edward, and Melissa Burkley. 2009. "Mythbusters: A Tool for Teaching Research Methods in Psychology." *Teaching of Psychology* 36 (3): 179–84.
- Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.
- Goldberg, Amie, Michael Russell, and Abigail Cook. 2003. "The Effect of Computers on Student Writing: A Meta-Analysis of Studies from 1992 to 2002." *Journal of Technology, Learning, and Assessment* 2 (1): 1–52.
- Jones, Byron, and Michael G. Kenward. 2014. *Design and Analysis of Cross-over Trials*. Third edition. Boca Raton: CRC Press/Taylor & Francis.
- Li, Tianjing, Tsung Yu, Barbara S. Hawkins, and Kay Dickersin. 2015. "Design, Analysis, and Reporting of Crossover Trials for Inclusion in a Meta-Analysis." *PLOS ONE* 10 (8): e0133023.
- Noyes, Jan M., and Kate J. Garland. 2008. "Computer- vs. Paper-Based Tasks: Are They Equivalent?" *Ergonomics* 51 (9): 1352–75.
- Rosen, Amanda M. 2018. "The Best Breakfast in Town: A Comprehensive Research Methods Project." *PS: Political Science & Politics* 51 (01): 173–77.
- Senn, Stephen. 2002. *Cross-over Trials in Clinical Research*. 2nd ed. Statistics in Practice. Chichester, (UK); New York: J. Wiley.
- Touchton, Michael. 2015. "Flipping the Classroom and Student Performance in Advanced Statistics: Evidence from a Quasi-Experiment." *Journal of Political Science Education* 11 (1): 28–44.
- Velasco, Marcela, and Gamze Çavdar. 2013. "Teaching Large Classes with Clickers: Results from a Teaching Experiment in Comparative Politics." *PS: Political Science & Politics* 46 (04): 823–29.

Endnotes

¹ Due to attrition, the total sample size was 53. See appendix A1 for a flow diagram tracking observations over the course of the experiment.

² Each exam was worth 20 points but, for easier interpretation of the results, we rescaled the scores to 100. See appendix A2 for additional details about the course and the experimental design.

³ A post-hoc power analysis of the repeated measures ANOVA with $\alpha=0.05$, $n=53$, and a correlation $r=0.5$ between exam scores, showed that our statistical test could detect a medium sized effect of $f=0.25$ with 0.94 Power ($1-\beta$). The effect we observed was $f=0.47$, which is considered to be large. Our test would have lacked sufficient power to find small effects ($f<0.2$), however.

⁴ See appendix for sample exam questions.

⁵ The exams analyzed in this paper were administered a month apart.

⁶ Field experiments are increasingly being used in classrooms to investigate the effectiveness of various technologies and teaching strategies (e.g. Velasco and Çavdar 2013).