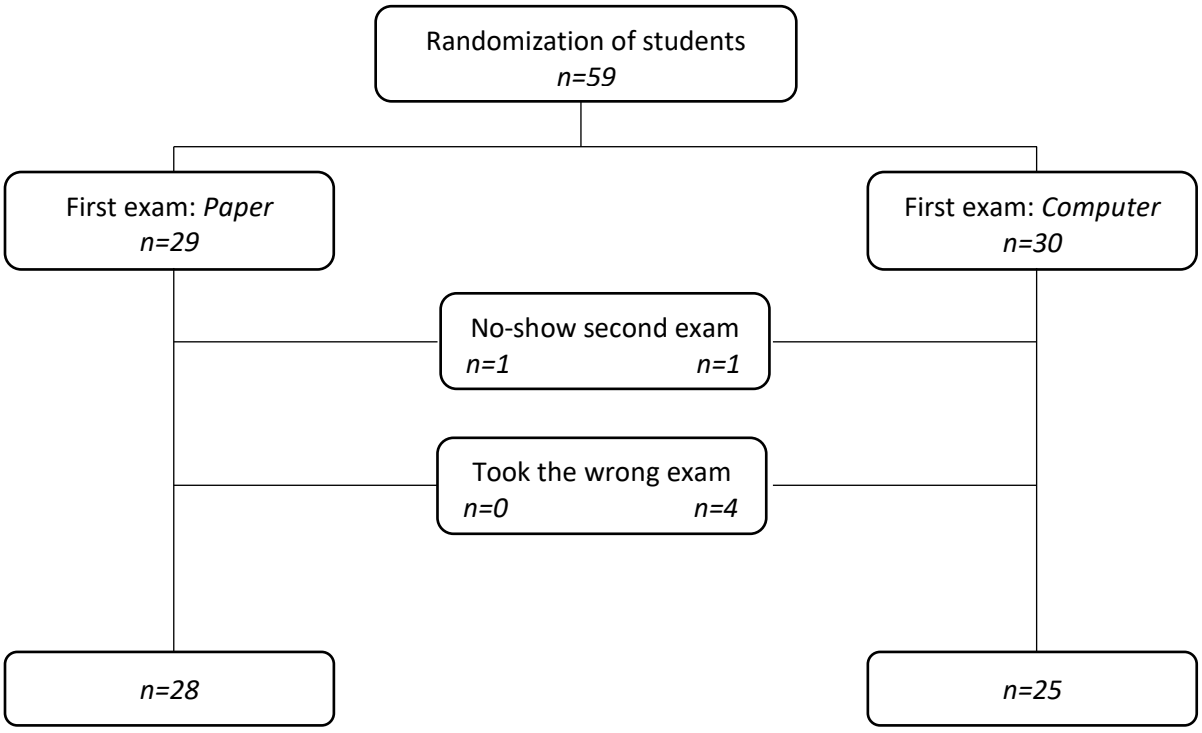
Appendix

A1: Flow diagram tracking observations over the course of the experiment



Note: The four students that took the wrong exam did so during the second exam by breaking their test-taking sequence. Specifically, all of them were supposed to take the second exam on paper but mistakenly took it on a computer. They were dropped from the analysis. The two no-shows for the second exam refer to students that dropped from the course. Five of our students had accommodations to take exams with extra time. We ran our analysis with and without these students. Since the substantive results did not change, we included all the students in the analyses discussed in the paper. The total sample size was 53.

A2: Additional information about the course and the research design

Course

In addition to the two exams used in the experiment, the students had to complete other course requirements including an additional paper-based in-class exam which tested them on research design and qualitative methods, two take-home exams, regular homework exercises and a final research project. To keep the instruments of our experiment as similar as possible, we used only the last two in-class exams that tested students on the quantitative part of the course to generate the data for the experiment. Together, the analysed exams were worth about 17% of the total grade.

Exams

The computer-based exams were administered on Moodle, which is our institution's course management software. To take the exams, the students had to log in on Moodle with their credentials. They could navigate back and forth between questions.

The paper-based exams were completed using "blue books," which the instructors provided to students at the beginning of the class together with the exam sheets. To minimize the waste of exam time, the "blue books" had the names of the students written on them and were placed on the desks of the first two rows of the classroom in an alphabetical order.

The questions on the exams tested student understanding of quantitative concepts and methods. Below are sample questions from each exam and a brief discussion of our goal for each followed by a sample student answer.

Exam 1:

1. What are the relative advantages and disadvantages of laboratory experiments as compared with experiments conducted outside the laboratory?

A good answer to this question discusses the principal strength of lab experiments (i.e. internal validity) but mentions limitations such as potentially low external validity, limited questions that can be studied etc. Furthermore, compared to lab experiments, field experiments have higher external validity but are more prone to inferential threats because of the inability to control for various factors.

Student answer:

Laboratory experiments have the advantage of allowing the researcher to manipulate only the desired variables and control the conditions. Because of this, lab experiments have a high amount of internal validity because the researcher can ensure that the experiment is unbiased. However, this high level of experimental manipulation can lead to less external validity, meaning that the results cannot be generalized outside the experiment. Experiments conducted outside the laboratory have high external validity because the observations reflect the true nature of individuals in an uncontrolled setting. But these experiments have lower internal validity because the researcher cannot control for external factors that could introduce bias.

2. The table below shows regression results where the dependent variable is Obama's vote share in the 2012 US Presidential election, and *Unemployment* is each state's November 2012 unemployment rate, with standard errors listed in parentheses.

    a. Interpret the coefficient for *Unemployment* and discuss whether it is statistically significant at the 0.05 level.
    b. How do you interpret the estimated intercept of this model?

	Obama 2012 Vote Share
	Model
Unemployment	2.142**
	(0.845)
Intercept	34.068***
	(6.112)
N	50
R-squared	0.118
Adj. R-squared	0.100
Residual Std. Error	9.930 (df = 48)
F Statistic	6.430** (df = 1; 48)

***p < .01; **p < .05; *p < .1
standard errors in parentheses

This question was intended to gauge basic understanding of OLS regression coefficients.

Student answer:

a) The coefficient for unemployment shows that for every one percent of unemployment that a state had, Obama won 2.142 more percent of the vote. Since it says that that this coefficient has a p value less that .05 we can say that it is statistically significant at the .05 level.

b) The intercept in this model means that Obama won 34% of the vote in states where there is no unemployment. While there in reality would be no states with zero unemployment, this intercept is important because it gives each state a baseline vote since there wouldn't be any states where no one votes for Obama regardless of employment.

Exam 2

1. You and some friends are arguing about whether family or money is the key to happiness. You turn to statistics to help inform your discussion. Specifically, you run a regression using the following model design:

   lm(happiness ~ number of kids + income)

   Using survey data, you measure "happiness" using a 0–10 point scale of reported happiness (where 0 is very unhappy and 10 is very happy); "number of kids", which reports the number of kids a respondent has; and "income", which reports a respondent's income in dollars. Estimating the regression, you get the following results:

   $$Y_i = 5.34 + 1.12(Number\ of\ Kids_i) + 0.087(Income_i)$$

   "See," your friend says. "The coefficient for income is much smaller than the coefficient for kids. That means kids are more important for happiness!" Respond.

This question was primarily designed to test whether students know how to interpret the relative substantive importance of regression coefficients and, secondarily, how to use regression results to make inferences. Below is good answer that demonstrates a mastery of both.

Student answer:

This statement cannot technically be verified or rejected because we do not have the data telling us whether or not either of the coefficients for income or number of kids is statistically significant. We also run into trouble trying to compare these two coefficients because the scale of the data is very very different. For every one increase in number of children, like the jump from 0 to 1 or 1 to 2, or coefficient tells us that there is an associated increase of 1.12 points of happiness. However, the scale for income is incredibly different from this. Most of the observations for number of kids won't grow much larger that 5 children, but the Income variable could go into the 100,000s. When we take the .087 coefficient and multiply it by 10,000 (for a

respondent's income in dollars), we would see that the influence that income has on happiness could potentially be much more important to this model.

2. In this problem, you will interpret the results of a logistic regression analysis of the relationship between gun control opinions and party identification. The binary dependent variable is coded 1 for pro-control opinions and 0 for anti-control opinions. The independent variables are "party ID," a 7-point scale, ranging from 0 (strong Republicans) to 6 (strong Democrats), and "female" a dummy variable coded 1 for females and 0 for males. Below are the logistic results:

Variable	Coefficient	Standard Error
Intercept	-1.999	0.32
Party ID	0.503	0.028
Female	0.689	0.595

Interpret the size and statistical significance of the regression coefficients.

The main goal of this question was to test whether the students understood how to interpret logistic regression output, especially using the shortcut "divide by 4" rule for the upper bound of the effect.

Student answer:

Since it is difficult to interpret coefficients of a logistic regression analysis, one uses the "divide by 4" rule to give an accurate representation of the data. One divides the regression coefficients by 4 in order to determine the largest probability increase in the dependent variable (the upper-bound probability) based on a one-unit increase in the independent variable. This occurs in the center of the probability distribution (0.5). The coefficient on Party ID (0.503) indicates that the upper-bound probability of the independent variable Party ID is 0.126. The coefficient is statistically significant since it is at least two standard errors from 0. The coefficient on Female (0.689) indicates the upper-bound probability of the independent variable Female is 0.172. This coefficient is not statistically significant, however, since it is not at least two standard errors from 0.

Connection between project and teaching: Bringing research methods closer to "students' sphere of interest"

One important feature of this research project, for us, was that we were able to use the experiment as a pedagogical tool to improve our instruction of research design and statistics. We designed our course to walk students through the different stages of a research project, from the generation of the initial question to the literature review and theory to operationalization and hypotheses to

design and finally data analysis and presentation. Essentially, our students were able to see – and participate in – a meaningful example of the execution of a research project from start to finish, while learning about the process in the abstract. Below we give some examples of how we used the experiment to foster learning in the classroom.

From the very start of the semester, we discussed the experiment with our students. We began by outlining how the study originated from a casual conversation with colleagues about the relative benefits of note-taking and test-taking on computers versus by hand. This was a nice opportunity to highlight the utility of social science research in systematically evaluating claims or received wisdom. We were also able to use the study to illuminate the process of asking good research questions and operationalizing key concepts. Our interest was in investigating whether the method of test-taking impacted student performance. Thus, when we asked our students to think of the most productive way to pose the central question that drove our experiment, students came up with the following analytical question:

- To what extent does the test-taking mode affect student performance?

After that, we operationalized performance in the different test-taking modes (i.e. scores from typed and handwritten exams), and formulated null and alternative hypotheses as follows:

- Null hypothesis: The test-taking mode has no effect on student performance.

- Alternative hypothesis: Taking tests on Moodle leads to systematically different performance than taking them on paper.

We then used this running example to emphasize the importance of devising an appropriate empirical strategy for hypothesis testing, which helped us introduce the field experiment and its main features. We ended the course with a presentation of the data analysis and the main findings. Walking students through this path allowed us to present them with a clear and compelling illustration of all the steps of the research process and explain why they are necessary for knowledge generation.

We were also able to use our experimental design to highlight many of the concepts important to research design and statistics. For instance, we were able to explain external validity with reference to our use of real-world test-taking processes and procedures. Further, concepts like internal validity and randomization were explained with reference to how our randomized 2x2 crossover design was ideal for accounting for between subject effects that might confound studies designed differently. These examples worked particularly well in that we were able to illustrate these important concepts by using a project that examined an outcome of direct relevance to our students. We found that our students were engaged throughout the semester and very eager to hear the results of the experiment once we had finished conducting the data analysis.

It is important to note that we conducted our study in a research methods and statistics class, and this provided us with a useful tool for instructing our students, but the utility of our design is not limited to these types of courses. Indeed, using student-generated data to examine questions of interest to students and teachers alike is a wonderful model for both engaging students in their own instruction and for addressing important questions in the science of teaching and learning.

Questions about note-taking preferences

We asked our students the following two questions: (1) Some students prefer to take notes for classes on computers while others prefer paper. When it comes to note-taking, would you say that you prefer: computers, paper, or indifferent between the two. (2) If you selected either computers or paper above, how strongly do you prefer your chosen method of note-taking (click not selected if you chose 'Indifferent'). The options ranged from weak preference (1) to extremely strong preference (5).