EXPERIMENTATION METHODOLOGIES FOR EDUCATIONAL RESEARCH WITH AN EMPHASIS ON THE TEACHING OF STATISTICS

by

Herle Marie McGowan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Senior Lecturer Brenda K. Gunderson, Co-Chair
Professor Vijayan N. Nair, Co-Chair
Professor Richard D. Gonzalez
Professor Edward A. Silver

To my Julia and my Isabel.

ACKNOWLEDGEMENTS

I owe a debt of gratitude to many people for helping through this dissertation process. Vijay, thank you for the ideas and guidance that made this dissertation possible. Brenda, thank you for your mentoring during my years at Michigan, and for always being willing to talk whenever I popped into your office. Oh, and thanks for the chocolate! Thank you to the members of my committee, Rich Gonzalez and Ed Silver, and to Joan Garfield for your valuable comments on my research. The final product is much improved due to your feedback. Thank you to all those who provided emotional and mental support throughout grad school: Stacey Culp, Lacey Gunter, Amy Wagaman, Carrie Hosman, Jane Kleyman, Joel Vaughan, Matt Linn, and Natallia Katenka. Special thanks to Joel, Carrie, Jane and Anindya for help with R code and with LaTeX.

Thank you to my family for your support throughout this entire process. Mom, I know you never stopped praying for me, and you never will. Your phone calls meant the world to me, even when I couldn't answer. Donna (MaMac), thank you for everything you have done, from cooking a month of meals while I was pregnant and studying probability theory, to housing my husband and kids for a month while I was desperately trying to finish (thanks to JoeMac for that last one, too).

And finally, completing this dissertation would not have been possible if not for my husband and daughters. Joshua, I would not have made it this far without your

shoulder to cry on and your willingness to discuss all things with me—from the state of statistics education to the price of housing at the beach. You are my best friend and my rock. Besos. Julia and Isabel, you will probably not remember the years I spent in school, but I always remember how your laughter helped me get though them. I hope this is an inspiration to you both, that you can do *anything* you set your mind to (with a lot of persistence and support from those who love you). I love you all more than I can say.

To God be the Glory.

I'm finished!

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

EXPERIMENTATION METHODOLOGIES FOR EDUCATIONAL RESEARCH
WITH AN EMPHASIS ON THE TEACHING OF STATISTICS

by

Herle Marie McGowan

Co-Chairs: Brenda K. Gunderson and Vijayan N. Nair

In this thesis, I explore the state of quantitative research in the field of statistics education. First, a content review from several prominent sources in statistics education is conducted. Based on this review, recommendations are made for advancing methodological research in this field.

Next, the design and analysis of a randomized experiment in an introductory statistics course are presented. In this experiment, factorial and crossover designs were used to explore several implementation aspects of "clickers", a technology for collecting and displaying, in real time, student responses to questions posed by the instructor during class. One goal was to determine which aspects were most effective in helping improve engagement and learning; another goal was to explore issues involved with implementing a large-scale experiment in an educational setting. The aspects explored were the number of questions asked, the way those questions were incorporated into the material, and whether clicker use was required or monitored. There was little evidence that clicker use increased engagement but some evidence that it improved learning, particularly when a low number of clicker questions were well incorporated into the material (vs. being asked consecutively).

Finally, a strategy for exploiting interactions between design factors and noise variables in the educational context is examined. The objectives of this strategy are: 1) Identify a teaching method that is robust to the effects of uncontrollable sources of variation on the outcome, or 2) Identify when a teaching method should be customized based on a noise variable. Achieving the first objective is desirable when there is heterogeneity in the noise variable within a class, for example, when the noise variable represents characteristics of the students themselves. The second objective involves using information in the interaction to proactively customize a teaching method to particular groups, and is easiest for noise variables measured at the instructor or classroom level.

CHAPTER I

Introduction

Over the past decade, there have been many changes in the way statistics is taught, several of which are described in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report [1]. Technology, from Power Point® to real-time simulations, has become increasingly pervasive in our classrooms. New learning methods, including more active learning, are now common to our pedagogical approach. Even the use of non-traditional assessment, such as data analysis projects, is becoming more popular. In addition to changes in *how* we teach, there have also been changes in *who* we teach. Enrollments in statistics courses are growing, bringing not only more students but also a more diverse student population, with a wide range of prior exposure to statistical content, majors, and expectations for learning in our course, among other things. The GAISE report focuses on practical recommendations for improving statistics instruction to help deal with such changes. While the report encourages continued use of technology, active learning, and alternative assessment, it recognizes that there is room for improvement in each of these areas. The GAISE recommendations help address what is probably the most prevalent challenge for the educator: How to best help students learn. Certainly this is a great concern for educators in any field, but there is an extensive body of literature that discuss why this is particularly difficult in statistics—namely that patterns of incorrect reasoning about concepts such as probability and variability abound and are very difficult to change [e.g. 42–44, 65, 99]. (Further discussion of these issues is not the goal of this thesis; instead the reader is referred to these references for in-depth coverage.) In addressing the issue of how to best help students learn, other challenges arise: What new technology or pedagogical approach should be used; how to assess if these new approaches are helping students learn more and/or are worth the time (and money). Well-designed research can be a powerful tool for addressing these challenges.

The recent report Using Statistics Effectively in Mathematics Education Research (SMER [4]) describes several goals of research and discusses the components of well-designed research programs that would be necessary to accomplish each goal. From this report, the goals of research are:

- *Generate* ideas: Identifying ideas and hypotheses to be studied.
- *Frame* research: Define constructs and measurement tools, consider logistics and feasibility.
- *Examine* research: Design and implement initial, small-scale studies to test hypotheses.
- *Generalize* research: Design and implement larger studies to confirm results in general population.
- *Extend* research: Design and implementation of follow-up studies to improve treatment or explore long-term effects.

Much of the current research in statistics education utilizes qualitative methodologies to *generate* and *frame* research questions that address the specific challenges in teaching statistics (see [44]). As more sophisticated research questions are being identified, we need to transition to the use of quantitative research methodologies to *examine* particular ideas or interventions. The ability to *generalize* and *extend* research depends not only on the actual success of the interventions, but also on the quality of the research methodologies used to investigate them. My research assesses the feasibility and effectiveness of applying quantitative methodologies in an educational setting to gain a better understanding of how to advance to the latter research goals stated in the SMER report.

In this thesis, I explore the state of quantitative research in the field of statistics education and make recommendations for advancing this field. In Chapter II, findings from a content review of relevant publications form several prominent sources in the field of statistics education are presented. Based on this review, recommendations are made for advancing methodological research in this field.

In Chapter III, findings from a randomized experiment in an introductory statistics course are presented. In this experiment, factorial and crossover designs were used to explore several implementation aspects of "clickers", a technology for collecting and displaying, in real time, student responses to questions posed by the instructor during class. One goal was to determine which aspects were most effective in helping improve engagement and learning; another goal was to explore issues involved with implementing a large-scale designed experiment in an educational setting. Finally, in Chapter IV, a strategy for exploiting interactions between design factors and noise variables in the educational context is examined. The objectives of this strategy are: 1) Identify a teaching method that is robust to the effects of uncontrollable sources of variation on the outcome, or 2) Identify when a teaching method should be customized based on a noise variable.

CHAPTER II

Critical Review of Research Methodology in Statistics Education

2.1 Introduction

Reviews of educational research in general, and statistics specifically, have been conducted before. Garfield and Ben-Zvi [44] focused on statistics education by reviewing research from diverse sources and fields such as psychology, mathematics education, and science education, looking primarily at the research question being asked. In contrast, others have conducted complete content reviews of research published in specific educational journals, such as the *American Education Research Journal*, the *Journal of Educational Psychology*, or *Research in Higher Education* [e.g. 36, 47, 48, 56, 63]. Many of these simply cataloged the analytic procedures used, recording the type and frequency of each procedure in order to gauge the level of statistical training an educational researcher would need to understand and evaluate the results of the published research. Exceptions to this include Hutchinson and Lovell [56], who cataloged aspects of study design and reporting in addition to analysis, and Kieffer, Reese and Thompson [63], who went beyond reporting a simple tally and examined if analysis and reporting in the reviewed studies were consistent with recommendations made by the Task Force on Statistical Inference [113]. This current review differs from these previous reviews in two important ways. First, all aspects of the methodological process are considered: The research question being investigated, the choices in study design, outcome considered, and analysis, and the issues in reporting of this process. Second, implications for conducting quantitative research in educational settings are discussed.

The goals of this chapter, therefore, are two-fold: 1) Examine current methodological practice in the field of statistics education by conducting a content review of prominent sources of research in this field and, 2) using this content review as a basis, discuss the challenges of conducting research in educational settings. Areas where further methodological research is needed to address these challenges are postulated throughout the discussion. This discussion has wider relevance to educational interventions conducted in disciplines other than statistics.

2.2 Methods

Three sources were considered for this review: the *Statistics Education Research Journal* (SERJ), the *Journal of Statistics Education* (JSE), and the *Proceedings of the International Conference on Teaching Statistics* (ICOTS). JSE and the proceedings of ICOTS were reviewed over the ten-year period from 1998 to 2007. SERJ was reviewed from the publication of its first volume in 2002 until 2007. These sources were selected because they pertain exclusively to statistics education and are prominent within this field. From these sources, individual studies were selected for review if they met two criteria:

1. There had to be at least one quantitative outcome considered. Studies that used both quantitative and qualitative methods were included in this review, but the qualitative methods were not reviewed in detail. Studies that exclusively used qualitative data collection and analysis techniques were not included.

2. There had to be inference about the success of an educational intervention. For the purposes of this review, an educational intervention is defined as an active change in curriculum, pedagogical approach, or use of technology in the classroom that is compared to some baseline or standard method of teaching. Studies could consider either paired comparisons (pre vs. post), or comparisons of two or more independent groups; they could consider just one change in standard practice (simple interventions; e.g. having students work in groups) or multiple simultaneous changes (complex interventions; e.g. group work and use of computer applets). Studies that were only descriptive or correlational in nature (for example, a study seeking to identify predictors of success in a course) were not included.

Papers that were not empirical (e.g. theoretical, expository, or editorial pieces) were excluded. From JSE, the "Teaching Bits" and "Datasets and Stories" sections were excluded from this review, as were the posters presented at ICOTS. These choices help maintain focus on the types of studies that could be used to *examine*, *generalize*, or *extend* a research question.

Thirty-two studies—six from SERJ, five from JSE, and twenty-one from the proceedings of ICOTS—met the criteria of being a quantitative educational intervention. For each study, the following characteristics were recorded (with the categories considered, where applicable):

- Student level (Elementary or Middle School; High School; Undergrad; Post-undergrad)
- Question asked (Use of technology; Non-technological new pedagogical approach; Other)
- Outcome (Attitudinal—Validated; Attitudinal—Not validated; Learning—Validated;

Learning—Not validated)

- Design (Randomized control trial; Paired (pre vs. post) design; Crossover (2 or more conditions); Observational—Case/Control; Observational—Matched; Other)
- Sample Size(s)
- Length of study (Less than full term; Full term; More than full term—Where "term" refers to the normal academic period for the student level/institution considered)
- Analytic technique(s) used (After observing which techniques had been used, this variable was categorized into: Analysis of variance; Regression; t-procedures for means; Other)
- Tool(s) used to deal with variation, including:
  - Blocking (whether is was used and, if so, what the specific blocking factors were)
  - Covariate adjustment (whether it was used and, if so, what the specific covariates were)
  - Random effects (whether they were used and, if so, what the specific effects were)
- Quality of reporting, including:
  - Which statistics were reported
  - If course or lecture descriptions were included
  - If baseline equivalence was addressed
  - If study attrition was addressed

For the categorized characteristics, Table 2.2 lists the number of studies reviewed that fell into each category.

Table 2.1: Summary of Research Study Characteristics for the 32 Educational Interventions Reviewed

Characteristic	Categories (# of Studies in Each Category)
Student level	Elementary or Middle School (7) High School (1) Undergrad (21) Post-undergrad (3)
Question asked	Use of technology (21) New pedagogical approach (9) Other (2)
Outcome	Attitudinal—Validated (7) Attitudinal—Not validated (2) Learning—Validated (0) Learning—Not validated (30)
Design	Randomized control trial (7) Paired: pre vs. post design (5) Crossover: 2 or more conditions (1) Observational - Case/Control (17) Observational - Matched (0) Other / Design not clear (2)
Length of study	Less than full term (11) Full term (19) More than full term (2)
Analytic technique	Analysis of variance (8) Regression (4) t-procedures for means (10) Other (8)
Tools to deal with variation	Blocking (16) Covariate adjustment (11) Random effects (8)

2.3 Findings

Findings from this content review are summarized and presented in six broad categories: Research questions, outcomes considered, design, analysis, tools to deal with variation, and issues in reporting. The appendix at the end of this chapter lists the papers reviewed. It is appropriate to mention here that specific examples are sometimes provided within each of the six categories considered, but the detailed findings of the reviewed studies are *not* discussed as to allow the focus to be on the research methodology.

2.3.1 Types of research questions asked

A review of the literature shows that use of technology is the hottest topic in research. With continual advancements in capability and decreases in cost, it is no wonder that educators are turning to technology in an effort to enhance teaching and learning. As the GAISE report notes, "technology has changed the way statisticians work and should change what and how we teach" [1, p. 12]. Research on technology has focused on using it to change *how* we teach. Several studies investigated changes in delivery systems for course content, either within the traditional classroom setting (for example, through use of video [e.g. 11]), or to replace the classroom entirely with online courses [e.g. 35]. Technology has been used to aid student understanding by illustrating difficult concepts (for example, using computer applets [e.g. 5]) or by reducing the need for hand calculation.

Most of studies reviewed asked the question: "Is this technology better than the standard way of teaching?" An important follow-up question should have been: "Why is this new technology better?" Technology is rapidly changing—new forms are always being developed and current forms are continually advancing in features and capabilities. For instructors, there are often large start-up costs to incorporating a technological advance into the classroom—with respect to both the financial investment in physical resources and the investment of time to learn a new technology or develop new classroom activities and assessments. Knowing that some form of technology improves student learning is of limited use once that technology is obsolete. We need to utilize methods that allow us determine the "active ingredient"—what particular aspect(s) of that technology is helping students learn—in order to recreate its success in future innovations. Clearly this would also be beneficial when considering educational interventions of a non-technological nature. Methods such as multifactor designs, which allow simultaneous investigation of several factors of interest, may be useful in distinguishing active components from inactive ones. However, they are rarely used in educational research (see Section 2.3.3).

Interestingly, no studies looked at technology to change *what* we teach. Cobb [3] has argued that wide-spread use of distribution-based tests—for example, the t-test—in the introductory statistics curriculum is a hold-over from the days of poor computing power. He advocates that randomization-based permutations tests—which he believes are more intuitive—could now be taught since computer power is no longer a concern. However, Kaplan [60] noted that current students may not have sufficient background in programming to implement these tests. Cobb's suggestion and Kaplan's concern could easily be transformed into research questions for future study—testing what effect the use of randomization tests has on student understanding, or exploring what computational skills/training students would need to successfully implement them.

Studies that did not focus on technology considered a diverse range of pedagogical practices. Several looked at active learning techniques (such as working in groups [e.g. 45]). Several explored the benefit of using particular approaches to develop statistical reasoning (for example, using concept maps [e.g. 17]). Interestingly, only one study investigated the effects of teacher training on student learning [61] and only one explored changes to curriculum [112]; perhaps this is reflective of a general lack of focus on statistics in primary and secondary school.

One point to be noted is that the changes studied were generally incremental rather than radical (e.g. adding or changing one component of a course, not restructuring the course or content completely; an exception to this would be studies of online learning). Small changes may be more practical to implement. Also, radical changes to a course may not be ethical for students' learning. However, if incremental changes in pedagogy are associated with incremental changes in "signal," learning or attitudinal effects may be difficult to detect. This is especially problematic given the numerous noise factors—related to student, instructor, or institution characteristics—that are present in educational data, and power that is restricted by classroom sample sizes. Large-scale, multifactor experiments could be used to test several treatments of interest without sacrificing power. Such designs are not without practical and ethical concerns, however; see Section 2.3.3 for a discussion on the use of multifactor designs in educational settings. And while many analytic techniques exist for reducing measured variation, a fundamental difficulty in educational research is that many important sources of variation are latent and cannot be measured directly (for example, student motivation to learn). The same is true for many of the outcome variables considered in educational research. Issues pertaining to measurement of latent variables involve a large body of research in and of themselves, spanning many disciplines. Educational researchers need to participate in this research by systematically identifying potentially important sources of variation that arise in educational settings and working to develop accurate measures of them.

2.3.2 Outcomes considered

Learning Outcomes

Nearly every reviewed study measured considered student learning as the primary outcome. Without exception, learning outcomes were measured using a non-validated measure, such as a course exam or activity. Use of course exams as an outcome is easy to implement and should result in "high quality" data (since all students have a vested interest in taking and trying their best on a course exam).

Unfortunately, reliance on course grades is problematic for several reasons. Courses differ with respect to topics covered, emphasis placed on each topic, and exam structure. An exam in one course may measure students' ability to reason statistically while an exam in another course may measure students' computational prowess. It follows that similar scores on different exams may reflect different levels of understanding. Also, to see if the results from one study are reproducible, researchers need to repeat the design, implementation, and assessment used as closely as possible. This cannot be done if the precise assessment instrument is not available.

Instead of course exams, researchers should use common, reliable and valid measures of student understanding. One newly published journal, *Technology Innovations in Statistics Education*, even states that papers using quantitative assessments should provide evidence of reliability and validity, and that "Student performance on a final exam or end of course grade would not generally pass these tests" (see `http://repositories.cdlib.org/uclastat/cts/tise/aimsandscope.html`). The Assessment Resource Tools for Statistical Thinking project (ARTIST; `https://app.gen.umn.edu/artist/`) has developed several instruments with demonstrated reliability and validity, including topic-specific scales and the Comprehensive Assessment of Outcomes for a first Statistics course (CAOS [31]), which could be used to measure students' conceptual understanding. However, widespread adoption of these instruments seems slow in coming—none of the studies reviewed here used them. These multiple-choice assessments do not involve any mathematical calculation, so it may be that educators do not want to use them in place of traditional course exams (which often do involve some calculations). An alternative is to use these assessments in addition to the standard final exam, but clearly this could lead to problems with lower student response rates or reduced data quality if many students do not take these assessments seriously. As an illustration of this, evidence of extensive guessing by students was found by researchers using an assessment that did not count towards students' course grades [83]. Perhaps a good compromise would be to include topic scale or CAOS questions as part of a course exam while also including problem solving involving calculation. Of course, care would need to be taken to ensure that such an exam is not too long.

There is an additional point of discussion here—namely that any assessment instrument is measuring student *performance*, perhaps more so than student *learning.* There are two issues with this: 1) Students may recreate or identify a correct answer without understanding why it is correct, and 2) students come into a course with varying levels of conceptual understanding, which affects our ability to detect learning that occurred during the course.

The first issue is difficult to deal with. Certainly course exams that focus on procedures will suffer greatly from this problem. An exam that focuses on the application and extension of concepts will be provide a more accurate measure of actual learning, but the format in which such an exam is presented may affect its ability to do so. For example, even a well-constructed multiple choice question (i.e. where each alternative represents a plausible answer) does not allow students to demonstrate their thought process and skilled test takers may be able to identify correct answers without understanding why they are correct. Questions that allow for an open-ended or essay-type response are the best format to allow students to demonstrate their understanding, but would be difficult to implement in large classes. Such an instrument would also be difficult to grade consistently, both within a class and across the range of classrooms that use it for research assessment. The college AP-Statistics course exam includes a section for open-response, but grading this exam is highly centralized and coordinated. Surely a set of validated open-response questions could be developed for use in statistics education research, but could such an exam be graded consistently across the various researchers who will use it? Also, students might be able to better recall (and circulate among their peers) a few essay questions (as opposed to the 40-multiple choice questions that comprise the CAOS exam), weakening the measure the longer it is used. Perhaps a compromise would be a series of short answer questions, which allow for free response but might be easier to score consistently. Future work could explore the feasibility of creating open-ended assessments that would be widely useful as research instruments.

The second issue is more mathematical in nature, and perhaps more concrete to deal with. We want to measure what students have learned above and beyond the knowledge they came into a course with. The use of difference scores $[Y_{post} - Y_{pre}]$ has been proposed as a solution to measure gains in knowledge, but are not without their problems. Consider, for example, those students with extremely high scores on a pretest. These students have little room for improvement and their scores will likely change on the posttest—even in the absence of any intervention—simply as a result of regression to the mean. Gain scores $[(Y_{post} - Y_{pre})/(\text{max score} - Y_{pre})]$ have been used in physics education to address this issue, but the use of both difference scores and gain scores remains controversial [see, for example 79, 80, 114]. An alternative solution to transforming the analyzed response could be to simply subset the sample data based on pretreatment scores. There are statistical trade-offs with this approach: Students with either extremely high or low scores are adding noise to the data so precision could be gained by removing their data from analysis, but this of course lowers the effective sample size and decreases power. Future research is needed to explore the use of each of these alternative to measuring learning, including identifying the circumstances under which each is (and is not) most effective. Attempts to resolve the controversy surround the use of difference and gain scores could more systematically explore the circumstances under which their use is appropriate [e.g. 114] or is not appropriate [e.g. 80]. Similarly, studies could explore the conditions under which the benefit of excluding extreme scores from analysis outweigh the costs.

Attitudinal Outcomes

Only two studies considered student attitudes as the primary outcome [7, 16], though several measured attitudes in addition to learning. When studies measured student attitudes, they typically used a reliable, validated instrument to assess attitudes. Several studies used the Survey of Attitudes Towards Statistics (SATS [98]), sometimes supplementing this with additional questions. While this is an improvement over the measurement of learning outcomes, problems with the measurement of attitudinal outcomes still exist. In particular, attitudes are often measured on ordered categorical (e.g. Likert) scales but little attention is paid to the variability that can arise through this measurement process. For example, one student's operational definition of "Likely" or "Unlikely" may differ from another student's definition, or a student's definition may change from the beginning to the end of term. Additionally, this data is often coded and analyzed as if it were truly numeric, ignoring the variability that exists in the distance between categories within a person or across different people. Future research is needed to develop methods that could quantify the sources and magnitude of variability that can arise when using ordered categorical scales. Perhaps something can be learned from the engineering literature on Gauge R&R (repeatability and reproducibility) studies. Gauge R&R is a technique used in industrial design to characterize the basic capability of a measurement system. Repeatability characterizes the within-instrument variability: When the same machine used by the same operator on the same part produces different measurements. Reproducibility characterizes the between-instrument variability: When the same machine used by multiple operators on the same part produces different measurements. On many survey instruments participants are required to map qualitative responses to numeric labels, such as rating their “agreement” with a statement on a 5-point scale. In these terms, repeatability characterizes the variation that could arise if the same person used different mappings each time they took the same survey (resulting perhaps from a change in mood or perception). Reproducibility characterizes the variation that could arise from different people each using different mappings when taking the same survey. Repeatability and reproducibility parallel the concept of reliability in the psychometric literature (they are distinct from the concept of validity, which pertains to bias rather than variability). Reliability is rarely reported and often misunderstood; few educational researchers recognize the need to calculate the reliability of an instrument or scale for each sample on which it is administered [52, 55]. Educational researchers need to pay careful attention to the variability that can arise through the measurement process, either through consideration of reliability or through the an adaptation of the principles of gauge R&R. In particular, it is important to characterize the sources and potential magnitude of such variation prior to using an instrument as they could overwhelm any treatment effects if not taken into account. One goal would be to develop methods to identify which sources of variation could be controlled for. Another goal would be to determine how many replications would be needed to detect a signal when averaging over uncontrollable variations.

2.3.3 Types of research designs used

It is well-know in the Statistics community that random assignment of individual participants to treatment groups is the best way to guarantee that those groups will be comparable prior to treatment. However, this can be difficult to do in educational settings. A handful of studies reviewed were able to randomly allocate individual students to treatment conditions. Still, in many of these cases there was a reliance on student volunteers to participate in the research, which could limit generalizability of the results. Randomization of individual students to different sections of the same college course may be difficult if those sections do not meet at the same time. Then researchers will have to content with students' scheduling constraints. At the elementary or high school levels, it may be easier to randomize individual students since they are all in school for the same hours.

When it is not possible to randomize individual students, entire groups (such as different sections of a large class) can be randomly assigned to treatment conditions. Group randomization cannot offer the same promise of baseline equivalence of groups as can individual randomization. These groups are often self-selected and compositional differences may exist between them. When there is only one group per treatment condition, treatment will be confounded with section (or instructor, day, time, etc). It would instead be better to randomize several groups to each treatment condition so that existing differences can be averaged over, but this would require extremely large class sizes or the accumulation of data over time. For example, one study accumulated a sample size of over 5,000 college students by repeating the treatment conditions over four semesters [54]. The majority of studies reviewed were not randomized on either the individual or group level, but were instead observational in design. When random assignment is based on existing groups or is not used at all, for ethical or logistical reasons, it is especially important that any pretreatment differences between groups be addressed. However, many studies failed to discuss pretreatment differences in their write-up or account for pretreatment differences in their analysis (see Sections 2.3.4 and 2.3.6 for further discussion).

Beyond the two-group comparison

Nearly all of the studies discussed in this review involved two-group comparisons of some new technology or teaching method to some "standard" teaching practice. Only a handful of reviewed studies compared more than two treatment groups—one study compared three groups (two new treatments to one standard treatment [37]) and four used a factorial design. These factorial experiments ranged from basic 2^2 designs (two factors of interest with two levels each, resulting in four possible treatment combinations) to 2^3 designs (three factors with two levels each, resulting in eight treatment combinations). While each of these was a full factorial—including one group for each combination of treatment factors—the studies with the larger 2^3 design used interesting methods to maximize their available power. For example, one administered the eight treatment combinations in a crossover fashion, where students experienced a different treatment combination each class period, instead of as separate groups (each treatment combination was replicated two times throughout the term) [72].

Perhaps the prevalence of simple comparisons among the studies reviewed relates to sample size and the corresponding considerations of time and money. Sample sizes in educational research are clearly limited by class sizes. Larger sample sizes can be accumulated by including multiple classes or schools. Larger sample sizes can also be accumulated over time, though this would obviously delay the results of the study and care would need to be taken to account for any time effects in the results. Limiting research to two-group comparisons could help make the most of available power, especially when considering the high level of noise that usually exists in human-subjects data. However, this limits the type of research questions that can be asked. As noted in Section 2.3.1, most studies asked the question “Is some new treatment (like a new technology) better than the standard way of teaching?” and that a necessary follow-up question is “Why is the new treatment better?” Being able to answer the second question allows us to discover what about that treatment is successful for helping students learn so that we can recreate this success in future treatments. Multifactor designs, like factorial and fractional-factorial designs, can be used for this purpose. They could also be used as screening experiments to explore complex educational interventions (those composed of several distinct treatments), then to refine and optimize important components of an intervention [see, for example, 27]. These designs can maximize available power while simultaneously investigating the effects (including interactions) of several treatments. Large, multi-section courses are becoming increasingly common in statistics. Additionally, there has been a recent focus on collaborative research in education [see 4]. Both of these could increase the feasibility of implementing such designs. However, they require a great deal of planning and coordination, especially to ensure that each course section has an equitable learning experience. Moreover, studying many treatments simultaneously may not be reasonable in the educational context. The practical issues of using multifactor designs in educational research need to be thoroughly explored. As a start, a case study of the design, implementation, and analysis of a 2-factor experiment in a large introductory statistical methods course is provided in Chapter 3.5.

2.3.4 Analytic techniques used

In the reviewed studies, the most common methods of analysis were analysis of variance and regression procedures. Other analytic techniques including paired t-tests [68, 76, 111, 112]; paired or independent z-procedures [25, 61, 105]; Wilcoxon signed rank tests [11, 88]; and Chi-square tests [103]. Interestingly, more than half of the observational studies reviewed used an analytic technique that did not account for any baseline differences between treatments groups. In any study, care should be taken—either through design or analysis—to reduce the biasing effects of pre-treatment differences. Thoughtful analysis is especially important in observational studies where no protection from bias is afforded via design. A few studies used nonparametric techniques, such

Perhaps the most striking feature discovered during this review is that the analytic methods used seem oversimplified given the complexity of the data being analyzed. The lack of attention paid to pretreatment differences between groups is one illustration of this. The use of groups—specifically group assignment to treatment and group delivery of treatment—leads to other fundamental complexities of educational data that needs to be considered when selecting a method of analysis. Many of the studies reviewed here used group assignment to treatment, but for most the analysis was conducted at the individual level. Even when individual assignment to treatment is used, given the very nature of educational research, treatment is delivered to a group; that is, all members of a group are exposed to the exact same treatment under the exact same conditions. So students are nested within classrooms which in turn are nested within schools. Student responses may be influenced by the mix of course, instructor, and school characteristics that become unintended but integral parts of their treatment experience. When student is the unit of analysis despite such use of groups, the assumption of independence between units—required for most statistical tests—is violated (see page 37 of the SMER report for a discussion of this). The grouped structure of educational data can be addressed through the use of hierarchical linear modeling [93], however only one study reviewed used this analytic approach [45]. Hierarchical modeling (also called multilevel modeling) will be especially important as the number of research collaborations—with data collected from many different classrooms that are each nested within different institutions—increases.

2.3.5 Tools to deal with variation

The tools used to deal with variation cover aspects of both design and analysis. In terms of design, blocking was used to reduce the influence of the blocking factor on the results. Blocking factors included site [61], instructor [e.g. 27, 46], textbook [e.g. 33, 101], prior student knowledge or ability [e.g. 30, 37], and time of day [54, 111].

In terms of analysis, covariate adjustment and random effects were each used to deal with variation. Most studies that adjusted for pretreatment covariates used measures of academic ability or knowledge that came from sources outside of the course itself, such as SAT/ACT scores or GPA [e.g. 101, 109]; pre-treatment measures of the outcome, such as a previous course exam, were used as covariates less often. Use of external measures of knowledge could reduce the burden of assessment (e.g. time alloted to complete an assessment, student anxiety in being assessed) during the course of the experiment. Most studies that used covariate adjustment only included one or two variables. Similarly most studies that used random effects typically only included an effect for student (though one study did include random effects for semester, instructor, class and lab section, as well as student [54]).

Given the myriad of potential sources of noise that can arise in educational settings—related to student, instructor, course, and school characteristics—dealing with variation should be of utmost concern to educational researchers. Standard techniques such as blocking, randomization, and covariate adjustment were used quite often in the studies reviewed. There is a data analysis strategy that could also be useful for dealing with uncontrollable sources of noise. This strategy involves exploiting interactions between design factors in a multifactor experiment and noise variables to achieve one of two objectives: 1) Identify a teaching method that is robust to the effects of uncontrollable sources of variation on the outcome (called robust design [106]), or 2) Identify when a teaching method should be customized based on a noise variable. Achieving the first objective is desirable when there is heterogeneity in the noise variable within a class, for example, when the noise variable represents characteristics of the students themselves. The second objective involves using information in the interaction to proactively customize a teaching method to particular groups, and is easiest for noise variables measured at the instructor or classroom level. In Chapter IV, application of this strategy is illustrated within the context of a hypothetical multifactor experiment in statistics education.

2.3.6 General issues in reporting

The studies reviewed varied greatly in the level of detail that was reported in the paper, with respect to both implementation and analysis of the research. There are many aspects to implementation to consider: Ideal implementation (what would be best to address the research question while minimizing bias and noise), planned implementation (given the constraints, what is the proposed design), and actual implementation (what was and was not accomplished). At the very least, actual implementation needs to be described in every research report. In the educational setting, this should include details about the course (e.g. topics covered; grading policies), students served (e.g. year in school, whether course fulfills a requirement for them), and instructor characteristics (e.g. years of experience). It should also include details pertaining to the treatment itself:

- What was the treatment? (e.g. use of software to illustrate concepts)
- How was the treatment used? (e.g. in guided laboratory sessions; on homework)
- Who used the treatment? / How was the treatment assigned? (e.g. self selection; only certain sections were given access)

Clearly the above is not an exhaustive list. The SMER Report [4] lists comprehensive reporting guidelines for various components of a research program. Perhaps the best all-encompassing guideline is to "provide enough information to allow replication of the study" (p. 18).

There was great variety in how studies chose to report on the existence of baseline differences between the treatment and control groups. A few addressed this explicitly through formal testing of group demographic variables; many more addressed it implicitly through the use of covariate adjustment. However, there were several studies that only mentioned differences could exist but did not report any pretreatment information on groups; several studies failed to discuss relevant differences at all. Similarly, many studies failed to discuss missing data or attrition in their report. Only one study reported results from a significance test to compare retention rates between the treatment and control groups, as well as interviews with students to find out why they had withdrawn from the class [111].

In terms of inferential analysis, most studies reported the value of the test-statistic, degrees of freedom, and p-value. Studies were not as consistent in the reporting of group means or standard deviations. Only two studies reported effect sizes or confidence intervals, which could be useful for gauging the magnitude of effects in addition to their significance. No study reported on the reliability and validity of their assessment instruments, or the soundness of assumptions that accompany their analyses.

The lack of reporting of some issues, such as baseline equivalence or inference assumptions, does not imply that such issues were not explored or analyzed—it's just that the corresponding information was not included in the paper. Of course, manuscripts often must meet page limits in order to be published, and there are practical limits to how much detail can be provided. Consensus should be reached within, and possibly across, statistics education journals as to what information is most important to include in each manuscript, perhaps with additional information available in an online appendix. This will make the body of research more transparent and enable replication of studies and comparison of results across many different settings.

2.4 Summary

Overall, thirty-two studies—six from SERJ, five from JSE, and twenty-one from the proceedings of ICOTS—met the criteria of being a quantitative educational intervention. The comparatively low number of quantitative studies in SERJ and JSE seems to reflect the general focus on qualitative research in statistics education, which centers on ideas such as identifying difficulties in learning or patterns of reasoning. In contrast, the International Conference on Teaching Statistics seems to attract more quantitative studies on the actual practice of teaching.

2.4.1 Summary of current research

A review of these thirty-two studies revealed several features about the current state of quantitative research in statistics education:

1. Statistics educators are very interested in the use of learning technologies to improve the teaching of Statistics.
2. Most researchers focus specifically on improving student learning, as measured through grades on course exams or activities.
3. Group assignment to treatment, which is often observational, and group delivery of treatment are common.
4. A wide variety of analytic methods are employed, the simplicity of which belie the complex nature of educational data.
5. To deal with expected sources of variation, most researchers use blocking, covariate adjustment, or a combination of both.
6. Reporting about study implementation and analysis is inconsistent, making it difficult to replicate studies in the future.

Discussion specific to many of these points was included in the corresponding subsections of Section 2.3.

The most striking overall conclusion of this review is the disjointed, ad-hoc nature of quantitative research in statistics education. Specifically, there does not appear to be any systematic approach to studying problems. Studies are conducted in isolation, without much connection to previous research—either qualitative or quantitative—limiting the gain and application of knowledge from such research. Garfield and Ben-Zvi [44] noted this in their review as well. However, it appears as though many researchers are interested in similar topics and could learn from the experiences of others. To address this, there has been a call for the development of collaborative research programs and the creation of groups to facilitate this collaboration (see, for example, the research arm of the Consortium for the Advancement of Undergraduate statistics education, `http://www.causeweb.org/research/`).

Recommended guidelines for future research into teaching and learning

As collaborative research programs are developed and the number of quantitative studies increases, it is important that consistent methodological guidelines be followed. Several recommendations can be made based on the findings of this review. Certainly these ideas are not new, nor are they exhaustive, but they are worth revisiting here.

- Follow-up questions of treatment efficacy with with questions that will allow for identification of the "active" ingredient(s) in the success of a treatment so that ingredient could possibly be replicated in future treatments.
- Use valid and reliable assessment instruments when measuring outcomes, particularly learning outcomes. Such outcomes already exist—the CAOS test, for example—and could easily be incorporated as part of existing course exams.
- Use multifactor designs to explore and refine complex treatments. This may be especially helpful for identifying active ingredients, as mentioned in the first recommendation. Further research is needed to explore the feasibility of implementing such designs (see Section 2.4.2).
- Take care in both the design and analysis of educational data to account for bias due to pretreatment differences, which could arise as a result of group assignment or group delivery of treatment.

- Use hierarchical modeling to analyze nested data. Given that nearly every educational intervention is implemented on groups of students nested within a classroom (and these classrooms are nested within schools, and schools within communities/cities, etc.), nearly every analysis in education should be hierarchical.

- Be detailed when describing the design and implementation of treatment and when reporting results such as means, standard deviations, test statistics, confidence intervals and/or effect sizes. Guidelines specifying the minimum amount of detail that needs to be reported should be developed and applied consistently across journals (see Section 2.4.2).

2.4.2 Summary of areas of need pertaining to methodological research

There are several areas of quantitative research methodology that need to be addressed with future research, which would enhance the recommendations above. These have been discussed throughout this chapter and are summarized here.

Pertaining to design, a systematic exploration of the feasibility of implementing multifactor designs in educational settings needs to conducted. These designs are uniquely suited to breaking down a treatment into parts in order to determine the "active ingredient" in the success of that treatment. In terms of analysis, future research could lead to improved measurement of ordered, categorical data (e.g. student attitudes measured on a Likert scale). There are existing statistical methods—like Gauge R&R studies—that could be used to quantify and account for the variability inherent in these scales. To better deal with the myriad of sources of variation that are present in education studies, there is an analytic strategy that could be used to exploit interactions between design factors and noise variables in order to design educational interventions that are either robust to uncontrollable variation or that have been customized to particular groups of students or instructors. Feasibility of applying this strategy in the educational context needs to be explored.

Finally, there needs to be a more consistent set of guidelines for the reporting of educational research. Enough information needs to be included in a published paper so that the results can be properly evaluated (within the context of the design and analysis) by readers and so that the study could be replicated. These guidelines could be developed through a survey of researchers to determine what information they believe is most pertinent to report, or perhaps the editors of the various statistics education journals could join together and develop guidelines based on their expertise and experience.

2.4.3 Final thoughts

Much of current research in statistic education is small and fragmented. Collaborative research programs are needed to systematically study the practices of teaching and learning statistics. The success of such research endeavors depends on the quality of each individual study conducted. Current quantitative research practices could be immediately improved by implementing some of the guidelines presented above. Quantitative research would be further improved through future contributions to the methodological challenges presented throughout this paper.

Appendix: Citations (in alphabetical order) for Sections 2.3.1 to 2.3.5

- **For Section 2.3.1: Types of research questions asked**
  - **Pertaining to technology**: Aberson et al. [5]; Alldredge and Brown [7]; Alldredge and Som [9]; Alldredge et al. [8]; Ayres and Way [11]; Enders and Diener-West [37]; Cicchitelli and Galmacci [25]; Collins and Mittage [27]; Davies et al. [30]; Dinov and Sanchez [33]; Dutton and Dutton [35]; Hilton and Christensen [54]; Lee [68]; Lipson [69]; Meyer and Lovett [76]; Meyer and Thille [77]; Stephenson [103]; Sundefeld et al. [105]; Utts et al. [109]; Ward [111]; Watson and Kelly [112]
  - **Non-technological pedagogical approach**: Bijker et al. [16]; Bolzan [17]; Enders and Diener-West [37]; Giambalvo et al. [45]; Ip [57]; Luchini et al. [70]; Mahmud and Robertson [72]; Periasamy [88]; Stangl et al. [101];
  - **Other research questions**: Kataoka et al. [61]; McLeod et al. [75]
- **For Section 2.3.2: Outcomes considered**
  - **Learning only**: Aberson et al. [5]; Ayres and Way [11]; Bolzan [17]; Enders and Diener-West [37]; Cicchitelli and Galmacci [25]; Collins and Mittage [27]; Davies et al. [30]; Giambalvo et al. [45]; Gonzalez et al. [46]; Ip [57]; Kataoka et al. [61]; Lee [68]; Lipson [69]; Luchini et al. [70]; Mahmud and Robertson [72]; McLeod et al. [75]; Meyer and Lovett [76]; Meyer and Thille [77]; Periasamy [88]; Stangl et al. [101]; Stephenson [103]; Sundefeld et al. [105]; Watson and Kelly [112]
  - **Attitudinal only**: Alldredge et al. [8]; Bijker et al. [16]
  - **Both learning and attitudinal**: Alldredge and Brown [7]; Alldredge

and Som [9]; Dinov and Sanchez [33]; Dutton and Dutton [35]; Hilton and Christensen [54]; Utts et al. [109]; Ward [111]

- **For Section 2.3.3: Types of research designs used**
  - **Randomized control trial—individual assignment**: Enders and Diener-West [37]; Cicchitelli and Galmacci [25]; Davies et al. [30]; Gonzalez et al. [46]; McLeod et al. [75]
  - **Randomized control trial—group assignment**: Alldredge et al. [8]; Hilton and Christensen [54]
  - **Observational case-control**: Aberson et al. [5]; Alldredge and Brown [7]; Alldredge and Som [9]; Ayres and Way [11]; Bijker et al. [16]; Bolzan [17]; Collins and Mittage [27]; Dinov and Sanchez [33]; Dutton and Dutton [35]; Ip [57]; Kataoka et al. [61]; Lipson [69]; Luchini et al. [70]; Stangl et al. [101]; Stephenson [103]; Utts et al. [109]; Ward [111]
  - **Paired (pre vs. post) or Crossover**: Lee [68]; Mahmud and Robertson [72]; Meyer and Lovett [76]; Periasamy [88]; Sundefeld et al. [105]; Watson and Kelly [112]
- **For Section 2.3.4: Analytic techniques used**
  - **Analysis of variance procedures—ANOVA, ANCOVA, MANOVA**: Aberson et al. [5]; Alldredge and Brown [7]; Alldredge and Som [9]; Alldredge et al. [8]; McLeod et al. [75]; Stangl et al. [101]; Utts et al. [109]; Ward [111]
  - **Regression**: Enders and Diener-West [37]; Collins and Mittage [27]; Dutton and Dutton [35]; Giambalvo et al. [45]; Gonzalez et al. [46]; Hilton and Christensen [54]; Mahmud and Robertson [72]; Stangl et al. [101]

  - **Independent-samples t-test**: Lipson [69]; Dinov and Sanchez [33]; Ip [57]; Bijker et al. [16]; Dutton and Dutton [35]
  - **Paired-samples t-test**: Lee [68]; Meyer and Lovett [76]; Utts et al. [109]; Ward [111]; Watson and Kelly [112]
  - **Other analytic methods**: Ayres and Way [11]; Bolzan [17]; Cicchitelli and Galmacci [25]; Kataoka et al. [61]; Luchini et al. [70]; Periasamy [88]; Stephenson [103]; Sundefeld et al. [105]

- **For Section 2.3.5: Tools to deal with variation**
  - **Blocking**: Aberson et al. [5]; Alldredge and Brown [7]; Alldredge and Som [9]; Alldredge et al. [8]; Enders and Diener-West [37]; Collins and Mittage [27]; Davies et al. [30]; Dinov and Sanchez [33]; Dutton and Dutton [35]; Gonzalez et al. [46]; Hilton and Christensen [54]; Kataoka et al. [61]; Lee [68]; Mahmud and Robertson [72]; Stangl et al. [101]; Stephenson [103]; Utts et al. [109]; Ward [111]
  - **Covariate adjustment**: Aberson et al. [5]; Alldredge and Brown [7]; Alldredge and Som [9]; Alldredge et al. [8]; Collins and Mittage [27]; Dutton and Dutton [35]; Hilton and Christensen [54]; Mahmud and Robertson [72]; McLeod et al. [75]; Stangl et al. [101]; Utts et al. [109]
  - **Random effects**: Enders and Diener-West [37]; Giambalvo et al. [45]; Hilton and Christensen [54]; Gonzalez et al. [46]

CHAPTER III

A Large-scale Designed Experiment Exploring the Effects of Clicker Use on Student Engagement and Learning

3.1 Introduction

As discussed in the previous chapter, the majority of quantitative studies in statistics education used 2-group, case-control designs to address the question: "*Is* some new technology/teaching method better than some standard approach?" It is important to follow up such a research question with an investigation into *why* or *how* a new method is successful. In this chapter, the experimental exploration of several aspects of "clicker"—a technology for collecting and displaying, in real time, student responses to questions posed by the instructor during class—is considered in an large introductory Statistics course.

3.1.1 A review of the literature on clickers

Clickers go by several names in the literature: Personal, student, audience, or classroom response systems are some of the most common. They have been used extensively in college courses, first and foremost in the field of Physics [e.g. 12, 29, 34, 38, 59]. They are also gaining attention in other fields, such as Medicine [e.g. 85, 90, 96, 108], Engineering [e.g. 32, 100, 118], Biology and Life Sciences [e.g. 18, 40, 91], Psychology [26, 81], Accounting [14, 23], Agriculture [28], Computer Science [62], Earth Science [49], and Statistics [94, 115]. More recently, clickers have been used in elementary and secondary education as well [24, 28, 51, 87].

Clickers have been used in the classroom for a variety of purposes. Predominantly, they are used to check students' understanding of a topic soon after it has been covered in class [e.g. 14, 49, 64, 66, 92]. They can also be used to check students' prior knowledge on a topic [12] or to see if assigned reading was completed [14]. Clickers can be used as a tool to gather data for analysis or case study [19, 26, 50, 53, 94], and they can facilitate the administration of quizzes or exams [21]. Clickers can be used to practice calculations or check understanding of vocabulary [49, 115], however many proponents of clickers suggest that the questions should be conceptual in nature rather than focus on procedures or memorization [e.g. 12, 29, 34]. For example, questions can be written to point out common misconceptions [12, 49], draw connections between distinct topics [12, 14], or distinguish between similar concepts [12]. Clickers can be used to stimulate classroom discussion by pointing out student perceptions of a situation or exploring the implications of an idea [12]. Discussion could also be stimulated through questions with multiple or subjective answers, however Greer and Heaney [49] found that this could frustrate students. Several guides for writing good conceptual clicker questions exist in the literature, including Beatty [12] (referenced several times in this paragraph already), Duncan [34], Beatty et al. [13], Zhu [117] and Caldwell [22].

Overwhelmingly, proponents of clickers cite two perceived strengths that could make them a valuable tool for education. First, clickers provide immediate feedback to both students and instructors during a lesson. Student responses to a question can be tallied in just a few seconds and displayed in bar-graph form, giving the instructor a chance to gauge the understanding of the class as a whole and students the ability to gauge their own personal understanding [20, 22, 32, 95]. Second, clickers may help students engage more fully with the material. Since individual responses are aggregated and displayed anonymously to the class (so that is it not possible to know which answer a particular student selected), students tend to feel more comfortable responding than if they had to offer a verbal answer [22, 58, 95]. Also, the interactive nature of clickers may help students pay more attention to each question [67, 78, 108]. Many students who have used clickers report that they improve the classroom experience [10, 67, 71, 107] and improve their own understanding of the material taught [10, 20, 22, 71, 91, 107, 108].

Unfortunately, empirical evidence to support student perceptions of increased engagement and learning has been mixed. In terms of engagement, few studies have gone beyond measurement via student report. Exceptions to this have focused specifically on student participation. For example Carnaghan and Webb [23] measured participation by counting the number of questions *asked* per student during lectures in which clickers were used as compared to lectures when clickers were not used. They found a significant decrease in the number of questions asked when clickers were used, perhaps because students are less likely to ask clarifying questions when they see a large proportion of their classmates answered correctly. VanDijk et al. [110] observed a similar decrease in questions asked by students when clickers were used, though they did not track this formally. On the other hand, Stowell and Nelson [104] measured participation as the number of questions *answered*—both formally, by responding to displayed multiple-choice review questions, and informally, by volunteering to answer an open-ended questions verbally posed by the instructor. They compared participation rates between three groups: one that used clickers, one that used lettered response cards, and one that simply raised their hands. They found no significant difference in informal participation rates between the three groups and found that formal participation was higher in the clicker and card sections than in the hand-raising section. Taken together these studies point to a potential trade-off when using clickers: students seem more comfortable responding to questions but less comfortable asking them.

In terms of learning, many studies have found higher exam scores when clickers were used [28, 40, 66, 85, 90, 91, 110]. It should be noted, though, that several of these studies demonstrated only conditional improvement. For example, Carnaghan and Webb [23] and Schackow et al. [96] found a significant improvement in scores only for those exam questions that were most closely related to the clicker questions asked during class. Kennedy and Cutts [62], Lass [66], and Nosek et al. [85] found that improved understanding was associated with increasing amounts of clicker use and/or better performance on clicker questions (i.e. answering more questions correctly). Unfortunately, analyses based on self-selected dose (i.e. student selected amount of clicker use) could be subject to selection bias, if it was the better students who chose to use clickers more and/or answered more questions correctly. Only one study formally manipulated the number of clicker questions asked during a semester: Preszler et al. [91] changed the number of questions asked in each lecture of several Biology courses between low (0-2 questions), medium (2-4 questions) or high (4-6 questions). They found a significant increase in exam scores as the number of clicker questions increased. There were also several studies that found no significant difference in exam scores for students who use clickers versus those who did not [73, 78, 96, 102, 104]. One study even found significantly worse exam scores for students using clickers: VanDijk et al. [110] compared three groups of students: 1) those in a traditional lecture section, 2) those in a clicker-only section, where questions were posed only once before an instructor-lead discussion of the answers and 3) those in a clicker section with Peer Instruction, where questions were posed twice with group discussion in between (see Mazur [74] for more on this). They found that students in the clicker-only group had lower exam score than students in the other two groups, which were similar in performance to each other. VanDijk et al. [110] attributed this lower performance to the fact that students in the clicker-only group seemed to ask fewer clarifying questions.

To add to the current understanding of clickers as a pedagogical tool—specifically to explore which features of clicker use might increase student engagement or learning—an experiment was conducted from January to April 2008 at the University of Michigan. This experiment took place in the laboratory sessions of a large, introductory data analysis course called Statistics 350: Introduction to Statistics and Data Analysis.

3.1.2 Description of Statistics 350

Statistics 350 is a 4-credit course taught every semester (14 week term) at the University of Michigan. Historically, most students taking this course are undergraduates who need to fulfill some graduation requirement, either for their major or the University in general. Course topics include descriptive statistics (numerical and graphical summaries), probability, sampling distributions, and inference procedures. The inference procedures covered include confidence intervals and hypothesis testing for proportions (one- and two-sample), means (one-sample, paired, independent, and one-way analysis of variance), simple linear regression, and chi-square analyses (goodness-of-fit, homogeneity, and independence). Students attend three hours of lecture and one 1.5 hour computer lab each week. The lecture sections vary in size, ranging from 60 students to over 400 students. The schedule of the lecture sections also varies: There are sections offered each week as three one-hour sessions, two ninety-minute sessions, and one three-hour session. For any given week, however, the same basic material is covered in all lecture sections. During the experimental semester there were six lecture sections taught by a team of four instructors.

Lab sections are more uniform than lecture sections in terms of size and structure; there are also many more lab sections, which allows for replication of treatment conditions. For these reasons, the experiment was implemented in the lab sections of the course. The goal of the labs is to reinforce concepts presented in lecture and provide hands-on examples of data analysis using the statistical analysis package SPSS. Occasionally some material is covered in lab before it has been presented in detail during lecture. The same activities—involving either computer-aided data analysis or solving word problems—are covered during each 90-minute lab under the guidance of a Graduate Student Instructor (GSI). During the experimental semester, there were 50 lab sections taught by a team of 24 GSIs (22 GSIs taught two sections each, two taught three sections each). The lab sections had a maximum enrollment of either 21 or 27 students, depending on classroom size. The sequence of topics covered during lab the semester were:

- **Lab 1** : Descriptive statistics and graphs
- **Lab 2** : Sequence and quantile-quantile (QQ) plots (not taught in lecture)
- **Lab 3** : Random variables
- **Lab 4** : Central Limit Theorem; Confidence intervals for a population proport
- **Lab 5** : Testing for a population proportion; Review for Exam 1

- **Lab 6** : Confidence intervals for a population mean
- **Lab 7** : One sample t procedures and paired t procedures for means
- **Lab 8** : Independent samples t procedures for the difference in means
- **Lab 9** : Independent samples z procedures for the difference in proportions
- **Lab 10** : One-way analysis of variance; Review for Exam 2
- **Lab 11** : Simple linear regression
- **Lab 12** : Chi-square analyses; Final Review

3.1.3 Previous clicker use in Statistics 350

Clickers have been used in Statistics 350 in a limited way since September 2006. From September to April 2006, clickers were used predominantly in labs and rarely in lectures. The TurningPoint® personal response system (`www.turningtechnologies.com`), which was recommended by the course textbook publisher, was used. With this system, students were only able to respond to multiple-choice questions. A set of the TurningPoint remotes was provided by the Statistics Department for use in the laboratory sections of the course, so that students did not need to purchase their own remote. This necessitated the time-consuming distribution and collection of the remotes during any lab period in which clickers were used. For this reason, they were only used in lab for three weeks of each semester to help students review for exams. Clickers were also used for a few lecture sessions per semester, but only in the smallest lecture section of the course (an evening section offered once per week with 60 out of approximately 1200 enrolled students).

Starting in May 2007, a new clicker system was introduced to Statistics 350, with the capability for students to input numeric responses as well as respond to multiple-choice questions. This new system—Qwizdom® (`www.qwizdom.com`)—was adopted as the official personal response system of the College of Literature, Science, and the Arts at The University of Michigan. The college provides technical support to both instructors who choose to use clickers and to students, who may use the same clicker remote for several classes throughout their college years. Students were required to purchase their own remote, which allowed clickers to be used in every lecture and lab session of the course. In lectures, clicker questions were typically interspersed throughout the material, often to view student responses to practice problems solved during class. Most lecture instructors asked the same questions, though they were not required to. In labs, a handful of clicker questions were usually asked at the end of lab, on the concepts that had been reviewed in that lab session. All GSIs were required to ask the same base set of questions but were allowed to add additional questions if desired, though this was rarely done. During the experimental semester, clicker use in lab was controlled (as described below) while clicker use in lecture continued as usual.

3.2 Design and Hypotheses

As stated previously, the goal of this experiment was to add to current understanding of clickers as a pedagogical tool—to explore which uses might increase engagement or learning. The specific research questions and hypothesis to be explored with this experiment were:

RQ1. Can you "overdose" on clickers by asking too many questions?

RQ2. What is the best way to distribute clicker questions throughout a class session?

H1. There will be a negative effect of clicker overuse—too many clicker questions asked consecutively.

RQ3. Are students motivated to use clickers even when it is neither required nor monitored?

The motivation for the first two research questions as well as the hypothesis came directly from my classroom experience. As a laboratory instructor when the TurningPoint system was first incorporated into Statistics 350, I noticed an increase in classroom disruption on the days clickers were used. The disruption occurred primarily while waiting for students to enter in their responses—some students responded quickly and became distracted (e.g. began talking or looking online) while waiting for the rest of the class to respond. This lead me to develop the belief that clicker overuse could actually be detrimental to students, particularly that there could be a negative interaction between the number of clicker questions asked and the way those questions are incorporated into the lesson. The motivation for the third research question arose from my review of the literature on clickers. Nearly every study about clickers reports that students perceived a benefit, in terms or increased engagement and/or learning, to clicker use. However, in each of these studies clicker use was required. The third research question seeks to see if students perceived benefit to using clickers will be enough to motivate their use when instructor-imposed incentives to do so are removed.

The primary outcomes of interest in this experiment are "engagement" and "learning." These terms are admittedly very broad in nature and difficult to measure. A review of the literature on engagement reveals that there are three aspects of engagement—behavioral, emotional, and cognitive [39]. Behavioral engagement involves doing the work and following the rules. Emotional engagement incorporates interest, values, and emotions. Cognitive engagement includes self-regulation, motivation, and effort. Studies with engagement as an outcome typically measure only the emotional aspect through student self-report of feelings and interest on a Likert-type scale. However, in this experiment all three aspects of engagement were considered (see Section 3.4). There has been a demonstrated link between engagement and learning, particularly when cognitive engagement takes place [39]. Student learning is typically defined as an improvement on a course-specific exam (e.g. a higher score on a posttest than on a pretest, or higher grades for one treatment group than another). As discussed in Section 2.3.2, one difficulty with the use of course exams to measure learning is that similar scores on different exams may in fact reflect different levels of understanding. To avoid such problems, several validated instruments, each from the Assessment Resource Tools for Improving Statistical Thinking project (ARTIST; `https://app.gen.umn.edu/artist/`), were used to measure student learning (see Section 3.4).

The treatment considered in this experiment is "clicker use." To define this more precisely, we focused on three specific components of clicker use which we believed might affect engagement and learning. These components, along with their measured levels, are:

1. Frequency: The number of clicker questions asked during a session
    (a) High: At least 6 clicker questions were asked
    (b) Low: 3-4 clicker questions were asked
2. Clustering: Asking all questions consecutively in a "cluster"
    (a) Off: Clicker questions were dispersed throughout the session
    (b) On: All clicker questions were asked consecutively, usually at the end of the session (operationally, a "cluster" was defined as 3 or more clicker questions in a row)

3. External Incentive: Whether clicker use was required, monitored, or not
   (a) High: Clicker use was required; student names were tracked using the clicker software and grades were assigned based on participation
   (b) Moderate: Clicker use was optional; student names tracked but no grades were assigned
   (c) Low: Clicker use was optional and anonymous; student names were not tracked (responses were saved under the anonymous heading "Participant i" for each student using clickers) nor were grades assigned

A 2 × 2 factorial design was used with the factors Frequency and Clustering to address the first two research questions and the hypothesis. The effects of Frequency and Clustering on emotional engagement, cognitive engagement and learning was explored. A crossover design was used to address the remaining research question. Specifically, the effect of External Incentives on behavioral engagement—namely, whether students choose to use the clickers when it was neither required nor monitored—was investigated. Since all students were required to purchase a clicker, all students were required to use their clicker at some point during the semester, so that no one felt their purchase had been unnecessary. For the *High* level of External Incentive, grades may be a powerful motivator to ensure that (most) students use the clickers. It should be noted, though, that grades were based on the student's general effort in answering clicker questions, not the number of questions they answered correctly. This was done primarily to reduce student anxiety about the questions; it has also been observed that grading based on effort ensures a more honest reflection of the class's level of understanding [59]. For the *Moderate* level, the incentive (or fear, as the case may be) of grading is removed, but the incentive of "we're watching you" remains. Students are perhaps so concerned about grades that even the potential to be graded may motivate them to use clickers. For the *Low* level, all external incentives have been removed—there is no way to even determine which students used the remotes. The belief is that if students perceive some value in the use of clickers—either that clickers make class more engaging or are helping them learn—then they will be motivated to use the clickers even as the level of external incentive decreases. In contrast, if students do not perceive real value in the use of clickers, they may not bother using the remotes when it is not required of them.

3.2.1 Assignment to treatment groups

During the semester in which the experiment was implemented, there were a total of 50 lab sections taught by 24 GSIs. Twenty-two of these GSIs taught two lab sections each, and two GSIs taught three lab sections each. Two separate randomizations were undertaken for the factorial and crossover designs. For the factorial design, the 24 GSIs were randomly assigned to one of four treatment groups and remained in this group for the entire semester. These treatment groups were identified by color for easy GSI reference. A summary of the design for the factorial experiment, along with the sample size for each group, is provided in Table 3.1.

Table 3.1: Factorial Design

		Clustering	
		On	Off
Frequency	Low	Team: Green n = 305 (93%)[a]	Team: Blue n = 279 (95%)
	High	Team: Orange n = 289 (93%)	Team: Yellow n = 324 (96%)

[a] n represents the number of students in each group who consented to have their data used in the experiment. The number in parentheses is the participation rate for that group—i.e. the percent of students assigned to the group who consented to have their data used.

For the crossover design, four crossover sequences were created based on possible combinations of the levels of External Incentive under the constraint that a switch between required (External Incentive = *High*) and optional (External Incentive = *Moderate* or *Low*) clicker use be made only once during the semester. The resulting sequences, along with the sample size for each, is presented in Table 3.2. The 24 GSIs were randomly assigned to one of the four sequences, independent of their randomization to the treatment groups of the factorial experiment. Within each sequence, GSIs remained at a given level for three weeks before switching to the next level in the sequence.

Table 3.2: Crossover Design

Sequence	Sample Size[a]
Low – Moderate – High	n = 297 (95%)
Moderate – Low – High	n = 287 (94%)
High – Low – Moderate	n = 306 (95%)
High – Moderate – Low	n = 307 (93%)

[a] n represents the number of students in each sequence who consented to have their data used. The number in parentheses is the participation rate for that sequence.

3.2.2 Correcting a limitation of previous studies on clicker use

One important aspect of the design of this experiment was to avoid confusion between the treatment of interest (roughly, "clicker use") and the simple pedagogical change of asking more interactive questions in class. This is a distinction that many studies on clickers have failed to make, so that results reported by these studies cannot be attributed to clickers themselves—it is possible that they are simply due to the practice of breaking up traditional lectures with questions [23]. A few studies did address this design flaw. For example, Schackow et al. [96] and Carnaghan and Webb [23] used crossover designs where students responded to multiple-choice questions verbally (presumably on a volunteer basis) or with clickers. Freeman et al. [40] compared two sections of a biology course; one section used clickers to respond to multiple-choice questions and the other used lettered cards to respond to the same questions. Similarly, my experiment was designed so that the exact same questions (with the same answer choices, when appropriate) were asked in every lab section. The sections differed with respect to the number of questions asked *using clickers*, the order of the clicker questions within the lesson (depending on whether or not those questions were clustered together) and the level of external incentive in encouraging students to use the clicker remotes.

3.3 Implementation Procedures

The experiment was conducted during the Winter 2008 term, which ran from January to April 2008 at the University of Michigan. The timeline of labs and the experimental procedures described here is given in Table 3.3. The treatment period did not begin until after the University's drop/add deadline, to ensure that class rosters were fixed (with the exception of a few students who dropped the course late). Prior to this, students experienced about three and a half weeks of lecture and three weeks of lab. Lecture topics covered during this pretreatment period included: descriptive statistics and graphs; sampling/gathering useful data; probability; random variables (binomial, uniform, normal); and inference for a single population proportion. Lab topics included: descriptive statistics and graphs; sequence and QQ-plots; and random variables.

A brief introduction to the experiment was provided to students during the first week of labs. Specifically, students were shown a slide with the following bulleted information:

- We believe using clickers will improve your learning experience, but are not sure of the best ways to use them.

Table 3.3: Experimental Timeline and Activities

Date	Lab Week	Activity for Experiment
		PRETREATMENT
January 3	–	None (1^{st} day of lectures; No labs)
January 7-9	1	Brief experiment introduction; Background information collected
January 14-16	2	None
January 21-23	–	None (No labs for MLK, Jr. Day)
January 23	–	NA (Drop/add deadline)
January 28-30	3	Formal experiment introduction; Informed Consent; First attitudes survey and CAOS
		TREATMENT PERIOD
February 4-6	4	Normal Distribution topic scale; Informed Consent for those absent from previous lab
February 11-13	5	(None other than clicker questions)
February 14	–	NA (Exam 1)
February 18-20	6	Sampling Distributions topic scale
February 22	–	Second CAOS due
February 25-27	–	None (Spring Break)
March 3-5	7	Confidence Intervals topic scale
March 10-12	8	(None other than clicker questions)
March 17-19	9	Hypothesis Testing topic scale
March 21	–	Third CAOS due
March 24-26	10	(None other than clicker questions)
March 27	–	NA (Exam 2)
March 31-April 2	11	(None other than clicker questions)
April 7-9	12	(Final attitudes survey and CAOS)
		POST-TREATMENT
April 15	–	None (Last day of lectures; No labs)
April 17	–	NA (Final Exam)

- So we will conduct an experiment with the clickers in labs this term, looking at
    - The number of questions asked in a session
    - How questions are incorporated into labs
- More info will come later . . .
- But don't worry—this will not mean any additional work outside of labs (unless it is for extra credit!)

At this point, students were asked to complete a background information survey. Note that while this was prior to completion of the formal informed consent process, it is common in the course for GSIs to collect similar information on their students to create example summary statistics and graphs.

There was no further mention of the experiment until the third week of labs, when students were given a formal description and asked to provide or refuse their consent to have their data used in our analyses. It should be noted that the entire assessment process, including the instruments selected and the manner in which they were administered, was designed to be an integral part of the course. This ensured that all students who were registered for the course after the drop/add deadline participated in experimental procedures—students provided consent only to allow their data to be analyzed. After the consent process, all students completed the pretreatment survey of attitudes towards Statistics and clickers as well as the pretreatment CAOS.

The treatment period began in the fourth week of labs. During the fourth week, students completed the ARTIST topic scale about the normal distribution. The other three topic scales were completed approximately every other week after that. As mentioned in Section 3.4, CAOS was completed around the time of each midterm exam in the course—at week six of the term and again at week nine. A final administration of CAOS and the attitudes survey took place during week 12. Throughout the treatment period (weeks four to twelve), several clicker questions were asked in each lab. The planned number of clicker questions for each week are presented by treatment group for the factorial experiment (Team) are provided in Table 3.4.

Table 3.4: Planned Number of Clicker Questions by Team

	Team[a]			
Week	Green	Blue	Orange	Yellow
4	3	3	6	6
5	4	4	11	11
6	3	3	7	6
7	4	4	6	7
8	4	3	9	9
9	4	3	11	8
10	4	3	10	10
11	3	3	8	8
12	3	3	6	6

[a] The teams are:
Green (Frequency=*Low*, Clustering=*On*);
Blue (*Low*, *Off*); Orange (*High*, *On*);
Yellow (*High*, *Off*).

3.3.1 Implementation infidelity

To better ensure consistency in teaching and grading among the 50 lab sections, there are weekly meetings for all GSIs to discuss what did or did not go well in the previous lesson, address questions about grading the homework, and to go over the lesson plan for the coming week. Every GSI gets a weekly memo with the meeting agenda as well as a schedule of specific activities to cover in the following lab. While consistency across labs is key, GSIs are still allowed freedom in how they will present the material (e.g. whether they create Power Point slides and how much content review to provide before starting the activities), allowing their own personality and teaching style to come though. During the experimental semester, it was necessary to reduce this freedom to some extent. For example, all GSIs were required to use the Power Point presentations with the questions for their treatment group (the questions were the same for each group, but the number and order of those asked with clickers varied) and were provided some guidance as to how to incorporate these questions into lab (e.g. to ask all questions at the end of the activity/lab or to incorporate the questions into the activity). However, restrictions on GSIs were kept to a minimum to avoid conflicts in the team or with the experimental procedure. In hindsight, the guidance provided as to the placement of clicker questions was not specific enough. GSIs varied in their interpretation of this guidance and their ultimate placement of the questions. It was not always clear to GSIs (especially those who were supposed to integrate questions throughout the lab material) when a question was to be asked *before* the corresponding material as opposed to *after*. This could affect the cognitive level of the question—a question which would have required deep thought before presentation of corresponding material simply requires recall ability when asked after.

During the experimental semester, the memos for the weekly meetings were personalized for each GSI, indicating their experimental conditions at the top of the page. This was to help them identify the appropriate presentation to use in lab—based on their assignment for the factorial experiment—and their crossover status for the week—based on their assignment for the crossover experiment. With this information, there were no GSIs who used the wrong team presentation. However, there were some discrepancies in the number of clicker questions assigned and the number actually asked due to technical or other issues in individual labs. Sometimes, due to technical or other issues, no questions could be asked with clickers. Considering all lab sections over the nine weeks of the treatment period, this occurred a total of 17 times (8 for labs assigned to High Frequency; 9 for labs assigned to Low). This discrepancy is perhaps more serious for those labs assigned to High Frequency, as they were essentially running at Low Frequency for those sessions. In addition to those times when no clicker questions could be asked, there were seven times when the High Frequency labs also ran under Low Frequency conditions (where the actual number of questions asked with clickers in a High Frequency lab was less than or equal to the planned number of clicker questions to be asked in Low Frequency labs). While the conditions of the crossover experiment were less subject to technical problems, there was confusion among GSIs that resulted in discrepancies between the assigned and the actual condition run. Recall that there were three crossover conditions—Low, Moderate, and High External Incentive, respectively—that were supposed to be run for three weeks at a time and then switched according to a randomly assigned sequence. While the condition to be used that week was included at the top of the memo, there were several GSIs who missed or did not understand this information. Two GSIs started the semester under the wrong condition; one of these realized their mistake and ran under the correct condition for the last week of the three week block (the other ran the incorrect condition for the entire three weeks). Seven GSIs did not make the switch properly at the end of the first three week block—six missed the switch and ran at their previous status and one switched to the wrong condition. In light of this, greater care was taken to emphasize the second switch the week before it was to take place. Still, one GSI missed the second switch and ran at their previous status for an additional week. Additionally, over the course of the experiment, GSIs reported that they forgot to announce their crossover condition to students about 6% of the time (when accounting for missing GSI reports regarding the announcement, the percentage could be as large as 17%), severely weakening any potential impact the External Incentive factor could have on student behavior.

3.4 Measures

Recall that three aspects of engagement—emotional, cognitive, and behavioral—were considered in this experiment. These were measured using various assessment instruments and methods, including:

1. Four subscales from the Survey of Attitudes Towards Statistics (SATS) [98]:
    - Affect: Positive and negative feelings about statistics
    - Value: Attitudes about the usefulness, relevance, and worth of statistics in personal and professional life
    - Cognitive competence: Attitudes about the intellectual knowledge and skills when applied to statistics
    - Effort: Amount of work expended to learn statistics

2. A survey on attitudes towards clickers, developed by the Center for Research on Learning and Teaching (CRLT) at the University of Michigan
3. By tracking the number of participants using clickers during each class

Questions from the SATS subscales were combined with questions from the CRLT survey on clickers to form a single assessment that was administered both prior to and at the conclusion of the treatment period. The Affect and Value subscales were used as measures of emotional engagement, while questions from the Cognitive Competence and Effort subscales of SATS were used as measures of cognitive engagement. Questions from the CRLT survey questions pertaining to clickers includes aspects of both emotional and cognitive engagement. Behavioral engagement was measured using the percent of students per lab section that used clickers under each level of External Incentive, where two levels of "clicker use" were considered: 1) Answering at least one clicker question; 2) Answering at least 50% of the clicker questions with the clicker remote during a given lab session. Note that is was not possible to track individual changes in clicker use across the three levels, as there was no way to identify individual students under the *Low* incentive level.

Learning was measured using several instruments from the ARTIST project:

1. Four topic scales:
    - Normal Distribution
    - Sampling Distributions
    - Confidence Intervals
    - Significance Tests
2. The Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) [31]

The ARTIST topic scales served as proximal measures of learning—the topic was covered in labs one week and then the corresponding topic scale was administered as soon after the corresponding topic had been introduced as the class schedule would allow (see Table 3.3 for the exact timeline). In contrast, CAOS is a comprehensive exam which meant to measure longer-term learning. CAOS was administered at four points throughout the experiment. The first administration took place during the third week of labs, the week before the treatment period officially began. The second and third times students completed CAOS outside of lab for extra credit; these took place during the sixth and ninth week of labs, respectively (around the week of each of two midterm exams). Finally, CAOS was completed during the last week of labs for the semester. CAOS served as both a pretreatment assessment of statistical knowledge and as a relatively distal, comprehensive measure of statistical knowledge. Each outcome measure—the SATS, CAOS, and ARTIST topic scales—was selected for use in this experiment because it is nationally available and has demonstrated content validity.

Measures of the planned treatment and the actual treatment received, where available, were also recorded. Indicators of the assigned treatment levels were coded as +1 for both the *High* level of Frequency and the *Off* level of Clustering, and -1 for both the *Low* level of Frequency and the *On* level of Clustering. Additionally, the actual number of clicker questions asked for each lab section was reported by the GSI each week. It was not possible, though, to collect specific details on the actual placement of each clicker question each week.

Finally, several student, lab, and GSI covariates were measured. Student background and demographic information included:

- Grade point average: Categorized as 1.7 to 2.6, 2.7 to 3.6, or 3.7 to 4.0

- Year in school: Freshman, Sophomore, Junior, or Senior
- Gender: 1 if male, 0 if female
- Lecture instructor: One, Two, Three, or Four
- Calculus experience: 1 if previously completed single- or multiple-variable calculus course, 0 otherwise
- Pre-calculus experience: 1 if previously completed pre-calculus or algebra course, 0 otherwise
- Credits: Number of other credit hours enrolled for during the term (not including the 4-credits for Statistics 350)
- Work: Average number of hours worked per week for pay (not on coursework) during the term

Lab and GSI characteristics included:

- Lab start time: Categorized as
  - Early morning: 8:30 am
  - Late morning: 10 or 11:30 am
  - Afternoon: 1, 2:30, or 4 pm
  - Evening: 5:30 or 8 pm
- Experience: Number of semesters the GSI had taught Statistics 350 prior to the start of the experimental semester

These particular variables were identified as potentially important covariates using two sources. Several variables were used for covariate adjustment in the studies reviewed in Chapter II. Additional variables were identified by the course instructors as sources of variation between students, labs or GSIs.

3.4.1 Sample characteristics

Any student who was registered for Statistics 350 after the drop/add deadline was eligible to participate in this study. Great care was taken to design an experiment that would fit seamlessly into the existing lab structure. As a result, all students were required to complete assessments for the experiment (or, in the case where assessments were optional, all students were given the same opportunity for extra credit). Therefore, students only had to provide consent for their data to be used for analysis. Students were assured that there would be no work above and beyond normal coursework and that their data would be confidential if they provided their consent. The overall consent rate was high—1197 (94%) of the 1277 enrolled students agreed have their data used in analyses.

Tables 3.4.1 and 3.4.1 have descriptive statistics for the covariates described above, for the entire sample (Overall) and by treatment group (Team). Some imbalances between the treatment groups can be seen. These likely result from the use of group randomization—students self-selected the lab section they wanted to attend and then entire sections were randomly assigned to treatment groups. The most notable imbalances are with the covariates Year and GSI Experience: The Blue Team has a disproportionately large number of Freshman and small number of Juniors and Seniors (see Table 3.4.1); additionally, the Yellow Team has a disproportionately larger average GSI experience (see Table 3.4.1). Covariate adjustment in analytic regression models can easily account for any discrepancies between the treatment groups due to the pretreatment covariates. Descriptive statistics show almost no imbalance for the covariate Credits and only minor imbalance for the variable Hours Worked, however, there is a large amount of missing data for Hours Worked (see Table 3.4.1). This suggests that little would be gained from including these two covariates in analytic models. For each analysis undertaken, model fitting will be employed to select relevant covariates for which to adjust (see Section 3.5).

Table 3.5: Summary of Student-level Covariates

		1.7 to 2.6	2.7 to 3.6	3.7 to 4.0		N
Grade Point Average	Overall	60 (5%)	748 (66%)	320 (28%)		1128
	Green[a]	14 (5%)	194 (67%)	82 (28%)		290
	Blue	12 (4%)	178 (66%)	79 (29%)		269
	Orange	17 (6%)	182 (69%)	64 (24%)		263
	Yellow	17 (6%)	194 (63%)	95 (31%)		306
		Freshman	Sophomore	Junior	Senior	N
Year	Overall	206 (17%)	473 (40%)	326 (27%)	189 (16%)	1194
	Green	42 (14%)	120 (39%)	98 (32%)	45 (15%)	305
	Blue	72 (26%)	116 (42%)	59 (21%)	31 (11%)	278
	Orange	37 (13%)	107 (37%)	93 (32%)	51 (18%)	288
	Yellow	55 (17%)	130 (40%)	76 (24%)	62 (19%)	323
		Female	Male			N
Gender	Overall	634 (57%)	473 (43%)			1107
	Green	167 (58%)	121 (42%)			288
	Blue	161 (60%)	107 (40%)			268
	Orange	160 (60%)	105 (40%)			265
	Yellow	146 (51%)	140 (49%)			286
		One	Two	Three	Four	
Instructor	Overall	412 (34%)	592 (49%)	53 (4%)	140 (12%)	1197
	Green	94 (31%)	164 (54%)	8 (3%)	39 (13%)	305
	Blue	100 (36%)	121 (43%)	9 (3%)	49 (18%)	279
	Orange	81 (28%)	163 (56%)	24 (8%)	21 (7%)	289
	Yellow	137 (42%)	144 (44%)	12 (4%)	31 (10%)	324
		Yes	No			N
Completed Calculus Course	Overall	881 (80%)	226 (20%)			1107
	Green	233 (81%)	55 (19%)			288
	Blue	213 (79%)	55 (21%)			268
	Orange	203 (77%)	62 (23%)			265
	Yellow	232 (81%)	54 (19%)			286
		Yes	No			N
Completed Pre-calculus Course	Overall	560 (51%)	547 (49%)			1107
	Green	153 (53%)	135 (47%)			288
	Blue	128 (48%)	140 (52%)			268
	Orange	144 (54%)	121 (46%)			265
	Yellow	135 (47%)	151 (53%)			286
		Min	Median	Mean (SD)	Max	N
Number of Credits	Overall	2	12	11.490 (1.980)	18	1113
	Green	3	12	11.440 (1.955)	18	286
	Blue	2	12	11.580 (1.918)	18	265
	Orange	3	12	11.450 (1.884)	17	262
	Yellow	3	12	11.500 (2.141)	18	300
		Min	Median	Mean (SD)	Max	N
Weekly Hours Worked	Overall	0	10	10.820 (8.133)	40	500
	Green	1	10	10.820 (7.689)	35	126
	Blue	0	10	11.500 (9.111)	40	113
	Orange	0	10	11.450 (8.338)	40	121
	Yellow	0	8	9.707 (7.445)	40	140

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

Table 3.6: Summary of Lab and GSI-level Covariates

		Early Morning	Late Morning	Afternoon	Evening	N
Lab Start Time	Overall	170 (14%)	320 (27%)	497 (42%)	210 (18%)	1197
	Green[a]	49 (16%)	121 (40%)	111 (36%)	24 (8%)	305
	Blue	72 (26%)	50 (18%)	95 (34%)	62 (22%)	279
	Orange	25 (9%)	74 (26%)	163 (56%)	27 (9%)	289
	Yellow	24 (7%)	75 (23%)	128 (40%)	97 (30%)	324
		Min	Median	Mean (SD)	Max	N
GSI Experience	Overall	0.0	1.0	2.2 (1.7)	6.0	1197
	Green	1.0	1.0	1.8 (1.0)	3.0	1197
	Blue	1.0	1.0	2.1 (1.2)	4.0	1197
	Orange	0.0	1.0	1.5 (2.1)	6.0	1197
	Yellow	1.0	3.0	3.1 (1.7)	6.0	1197

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

3.5 Analysis of the Experiment

This section presents analyses of all outcomes considered for the factorial and the crossover experiment. Outcomes pertaining to engagement are presented first, followed by outcomes pertaining to learning. For each analysis presented, the assigned treatment, rather than the treatment actually received, was analyzed to avoid bias in the estimated effects that could result from infidelity in the treatment implementation.

3.5.1 Emotional and cognitive engagement outcomes: The Survey of Attitudes Toward Clickers and Statistics

Recall that statements on the attitude survey were drawn from the Survey of Attitudes Towards Statistics (SATS) [98], as well as a survey on attitudes towards clickers developed by the Center for Research on Learning and Teaching (CRLT) at the University of Michigan. The Affect and Value subscales of the SATS were used as measures of emotional engagement. Statements from the Cognitive Competence and Effort subscales of the SATS were used as measures of cognitive engagement. Statements from the CRLT survey pertaining to clickers included aspects of both emotional and cognitive engagement. Students rated their agreement with each statement on a 5-point Likert scale ranging from Strongly Disagree (1) to Strongly

Agree (5), with a rating of "3" indicated neutrality ("Neither agree nor disagree"). statements that were negatively worded were reverse coded for the analyses.

Students completed the entire attitudinal survey both before and after the treatment period. Table 3.5.1 presents descriptive statistics, including Cronbach's α, of the pretreatment mean ratings for each of the five subscales for the entire sample (Overall) as well as by treatment group (Team). Table 3.5.1 presents the same information for the post treatment average ratings. Cronbach's α [86] is a measure of the reliability of the attitude ratings for this sample. Values range between 0 and 1, with higher values indicating better reliability. It is commonly held that values of $\alpha \geq 0.70$ demonstrate acceptable reliability. With the exception of the pretreatment Effort subscale, the values of Cronbach's α for this data are indeed high. Students were apparently not very consistent in their initial responses to the four items on Effort subscale, but these reliabilities improve to reasonable levels on the post treatment survey. Interestingly, the average of the mean ratings is largest for the Effort subscale at both timepoints, while average ratings were lowest for the Affect subscale both before and after treatment. For all scales, there appears to be a slight decrease in the average of the mean ratings from pre to post treatment. Similar decreases have been observed using the SATS before [97]. Also, it is possible that this decrease was influenced by grades on the course midterms: Students had received their scores on the second midterm (which are typically lower than scores on the first midterm; during the experimental semester, the average score decreased by three points from midterm one to midterm two) the week before completing the post treatment survey.

Table 3.7: Descriptive Statistics for Average Ratings on the Pretreatment Attitude Survey

	Team[a]	Cronbach's α	Min	Median	Mean (SD)	Max	N
Affect (Mean of 6 Statements)	Overall	0.82	1.00	3.50	3.42 (0.72)	5.00	1160
	Green	0.84	1.00	3.50	3.43 (0.73)	5.00	1148
	Blue	0.83	1.00	3.50	3.44 (0.73)	5.00	1149
	Orange	0.82	1.33	3.50	3.41 (0.73)	5.00	1157
	Yellow	0.80	1.17	3.50	3.40 (0.69)	5.00	1149
Value (Mean of 9 Statements)	Overall	0.86	1.89	3.78	3.80 (0.56)	5.00	1157
	Green	0.86	2.11	3.78	3.76 (0.58)	5.00	1144
	Blue	0.84	1.89	3.89	3.87 (0.52)	5.00	1147
	Orange	0.86	2.00	3.78	3.75 (0.58)	5.00	1152
	Yellow	0.86	2.00	3.78	3.80 (0.56)	5.00	1145
Cognitive Competence (Mean of 6 Statements)	Overall	0.85	1.17	3.83	3.76 (0.66)	5.00	1155
	Green	0.86	1.17	3.83	3.79 (0.69)	5.00	1141
	Blue	0.85	1.83	3.83	3.77 (0.66)	5.00	1144
	Orange	0.82	2.00	3.83	3.80 (0.63)	5.00	1150
	Yellow	0.84	1.33	3.83	3.70 (0.67)	5.00	1143
Effort (Mean of 4 Statements)	Overall	0.49	1.75	4.50	4.40 (0.52)	5.00	1163
	Green	0.46	2.00	4.50	4.38 (0.51)	5.00	1153
	Blue	0.43	2.00	4.50	4.46 (0.50)	5.00	1152
	Orange	0.57	1.75	4.50	4.40 (0.53)	5.00	1159
	Yellow	0.46	2.00	4.25	4.35 (0.53)	5.00	1154
Clickers (Mean of 12 Statements)	Overall	0.90	1.00	3.75	3.67 (0.62)	5.00	1136
	Green	0.90	1.00	3.75	3.66 (0.61)	5.00	1118
	Blue	0.89	2.08	3.75	3.72 (0.58)	5.00	1120
	Orange	0.91	1.17	3.75	3.67 (0.64)	5.00	1128
	Yellow	0.90	1.08	3.67	3.62 (0.63)	5.00	1117

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

Emotional Engagement

Figure 3.1 plots the average of the mean post treatment ratings by treatment factor for the Affect and Value subscales of the SATS, used to measure emotional engagement. In both plots, there appears to be an interaction. For the Affect subscale, this interaction is qualitative—that the *On* level of Clustering appears better than *Off* when Frequency is *High*, but not when Frequency is *Low*. In contrast, for the Value subscale, the *Off* level of Clustering is always better than *On*, with the difference being larger for the *Low* level of Frequency. However, the magnitude of the differences between the team averages for each scale are extremely small. To test if there is a significant effect of Frequency and Clustering on emotional engagement, two hierarchical linear models (HLM) were fit including nested random effects for GSI and lab section. For the first model, the response was the average rating on the Affect subscale; for the second, the response was based on the Value subscale.

Table 3.8: Descriptive Statistics for Average Ratings on the Post treatment Attitude Survey

	Team[a]	Cronbach's α	Min	Median	Mean (SD)	Max	N
Affect (Mean of 6 Statements)	Overall	0.83	1.00	3.50	3.37 (0.77)	5.00	1118
	Green	0.84	1.00	3.50	3.35 (0.78)	5.00	1100
	Blue	0.84	1.00	3.50	3.38 (0.78)	5.00	1105
	Orange	0.84	1.00	3.50	3.41 (0.79)	5.00	1091
	Yellow	0.82	1.00	3.50	3.34 (0.74)	5.00	1097
Value (Mean of 9 Statements)	Overall	0.86	1.00	3.67	3.66 (0.62)	5.00	1105
	Green	0.86	1.00	3.67	3.63 (0.61)	5.00	1085
	Blue	0.86	2.22	3.78	3.73 (0.57)	5.00	1088
	Orange	0.89	1.78	3.78	3.65 (0.66)	5.00	1074
	Yellow	0.85	1.89	3.67	3.66 (0.62)	5.00	1081
Cognitive Competence (Mean of 6 Statements)	Overall	0.83	1.17	3.67	3.63 (0.69)	5.00	1116
	Green	0.82	1.17	3.83	3.63 (0.67)	5.00	1100
	Blue	0.83	1.33	3.67	3.65 (0.71)	5.00	1102
	Orange	0.83	1.67	3.83	3.67 (0.70)	5.00	1089
	Yellow	0.81	1.17	3.67	3.56 (0.67)	5.00	1092
Effort (Mean of 4 Statements)	Overall	0.88	1.00	4.25	4.05 (0.74)	5.00	1122
	Green	0.94	1.00	4.25	4.06 (0.77)	5.00	1104
	Blue	0.83	1.00	4.25	4.06 (0.72)	5.00	1110
	Orange	0.89	1.25	4.25	4.03 (0.76)	5.00	1095
	Yellow	0.82	1.50	4.00	4.04 (0.71)	5.00	1104
Clickers (Mean of 12 Statements)	Overall	0.92	1.08	3.75	3.63 (0.69)	5.00	1101
	Green	0.91	1.25	3.75	3.61 (0.68)	4.92	1081
	Blue	0.92	1.17	3.75	3.62 (0.72)	4.92	1088
	Orange	0.92	1.33	3.75	3.63 (0.69)	5.00	1068
	Yellow	0.92	1.08	3.83	3.66 (0.69)	5.00	1071

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

Use of hierarchical modeling is necessary throughout the analyses here to account for complexities in the design. Specifically, students were nested within a lab section, lab sections were nested within a GSI, and GSI was the unit of random assignment. Additionally, to account for baseline differences between treatment groups that could exist due to the use of group randomization, several potential confounding variables were considered for inclusion in this model. A two-step backward selection procedure was used to identify important covariates:

1. All covariates described in Section 3.4 were initially included in the model. (Recall from the discussion in this section that the variables Credits and Hours worked were not considered for inclusion in the full model.) Covariates were included in the model in their order of believed importance (i.e. the variables believed to have the largest potential to impact the analysis were included in the model first; those believed to have smaller potential to impact the analysis

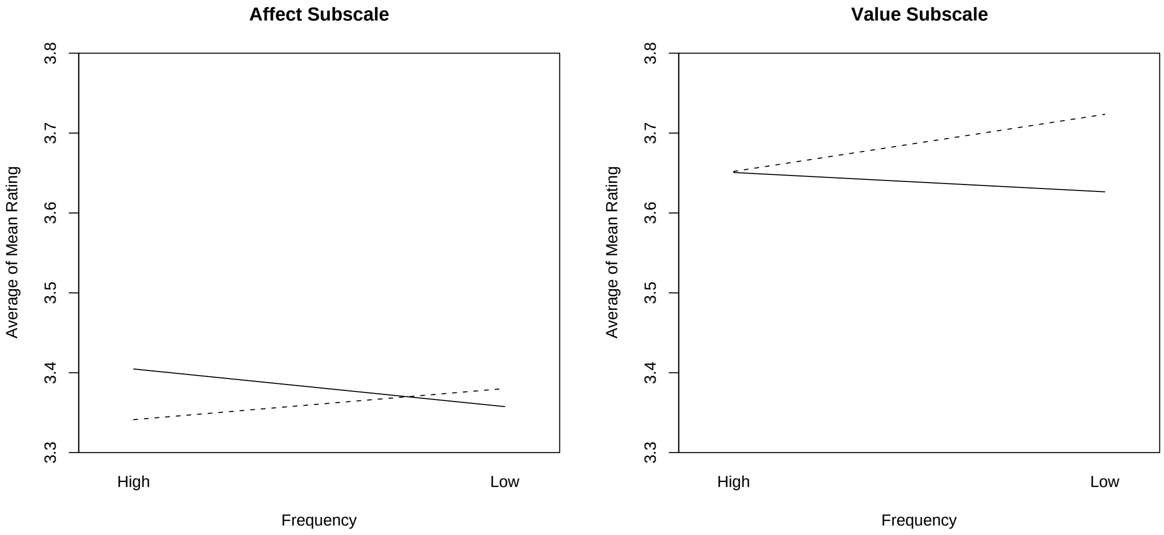

Figure 3.1: Average Mean Post Treatment Ratings by Design Factor for Scales Measuring Emotional Engagement. In each plot, the solid line corresponds to Clustering *On*, the dashed line to Clustering *Off*. Both plots are scaled to have the same range of 0.2 points on the y-axis.

were included last). Covariates that were insignificant at the 10% level were individually dropped from the model until only significant covariates remained, subject to the following constraints:

- The pretreatment measures of statistical knowledge (percent correct on the first CAOS) and general attitude toward statistics and clickers (average rating from the entire pretreatment attitudinal survey) could not be dropped. Note that these covariates were each centered at their respective overall mean.
- Levels of categorical variables were dummy (0/1) coded so that the largest category was the reference group. If one level of a categorical variable was significant, the entire variable was included in the model.
- Indicators of the crossover sequence that a particular GSI had been randomized to could not be dropped. These were included in the model to account

for any effects of the treatment factor External Incentive, which were not of particular interest when estimating the effects of Frequency and Clustering but needed to be accounted for.

- The main effects and interaction of Frequency and Clustering could not be dropped. To establish statistical significance of these effects, a 5% significance level used for the main effects and a 10% level was used for the interaction. (Also, recall that these effects were coded as -1/+1, not as 0/1.)

2. After all non-significant covariates were removed, the Akaike information criterion (AIC) [6] for the reduced model was compared to the AIC for the full model, and the one with the smaller AIC was taken as the final model. AIC is a tool for model comparison and selection that tries to balance model fit with the number of parameters. The model with the lowest AIC has the best fit for the smallest number of parameters.

Table 3.9 shows results for the hierarchical models for the Affect and Value subscales resulting from this selection procedure. For the model of students' affective feelings toward statistics, the estimated effects were 0.045 points for Frequency, -0.071 points for Clustering, and -0.003 points for their interaction (see top half of Table 3.9). The estimated variance components for this model were $\hat{\sigma}_{gsi} = 0.062$, $\hat{\sigma}_{lab} \approx 0$, and $\hat{\sigma}_{\epsilon} = 0.619$. For the model of students' value of statistics, the estimated effects were 0.017 points for Frequency, 0.004 points for Clustering, and 0.020 points for their interaction (see bottom half of Table 3.9). The estimated variance components for this model were $\hat{\sigma}_{gsi} = 0.019$, $\hat{\sigma}_{lab} = 0.027$, and $\hat{\sigma}_{\epsilon} = 0.514$. In both cases, the estimated effects are non-significant and each of the estimated variance components is small. The largest relative contribution to variation in each model is not surprisingly due to residual factors, including differences between individual students.

Table 3.9: HLM Results for Subscales Measuring Emotional Engagement

Affect Subscale				
	Estimate	Std.Error	DF	P-value
Intercept	3.157	0.074	877	0.000
Pretreatment CAOS	0.012	0.002	877	0.000
Pretreatment Attitudes	0.913	0.053	877	0.000
Year: Freshman	-0.037	0.060	877	0.533
Year: Junior	-0.121	0.052	877	0.021
Year: Senior	0.070	0.063	877	0.269
Calculus	0.300	0.054	877	0.000
Crossover Sequence 2	0.072	0.076	17	0.353
Crossover Sequence 3	0.051	0.074	17	0.506
Crossover Sequence 4	0.020	0.072	17	0.785
Frequency	0.045	0.025	17	0.383
Clustering	-0.071	0.028	17	0.225
Interaction	-0.003	0.026	17	0.949

Value Subscale				
	Estimate	Std.Error	DF	P-value
Intercept	3.577	0.062	863	0.000
Pretreatment CAOS	0.006	0.002	863	0.000
Pretreatment Attitudes	0.773	0.044	863	0.000
Grade Point Average: Low	0.027	0.086	863	0.757
Grade Point Average: High	0.074	0.039	863	0.058
Year: Freshman	-0.052	0.049	863	0.292
Year: Junior	-0.002	0.043	863	0.971
Year: Senior	0.194	0.053	863	0.000
Gender: Male	-0.169	0.035	863	0.000
Calculus	0.076	0.045	863	0.089
Crossover Sequence 2	0.204	0.057	17	0.002
Crossover Sequence 3	0.063	0.055	17	0.269
Crossover Sequence 4	0.037	0.054	17	0.498
Frequency	0.017	0.018	17	0.658
Clustering	0.004	0.021	17	0.923
Interaction	0.020	0.019	17	0.608

Note: Estimates reported for Frequency, Clustering, and the Interaction reflect the coding of these factors. That is, since these factors were coded as -1/+1, the estimated regression coefficient was multiplied by two to find the effect of going from the lower level of the factor to the higher level.

Cognitive Engagement

Figure 3.2 plots the mean post treatment ratings by design factor for the Cognitive Competence and Effort subscales of the SATS, used to measure cognitive engagement. As with the subscales measuring emotional engagement, there appears to be an interaction in both plots, though it is slight for the Effect subscale. In fact, the magnitude of the differences between means for the Effort subscale are nearly zero. For the Cognitive Competence subscale, *On* level of Cluster actually appears better than *Off* for the *High* level of Frequency and no worse than *Off* for the *Low* level of Frequency. Here again the differences in means is small, indicating that there may not be a significant difference between the treatment groups.

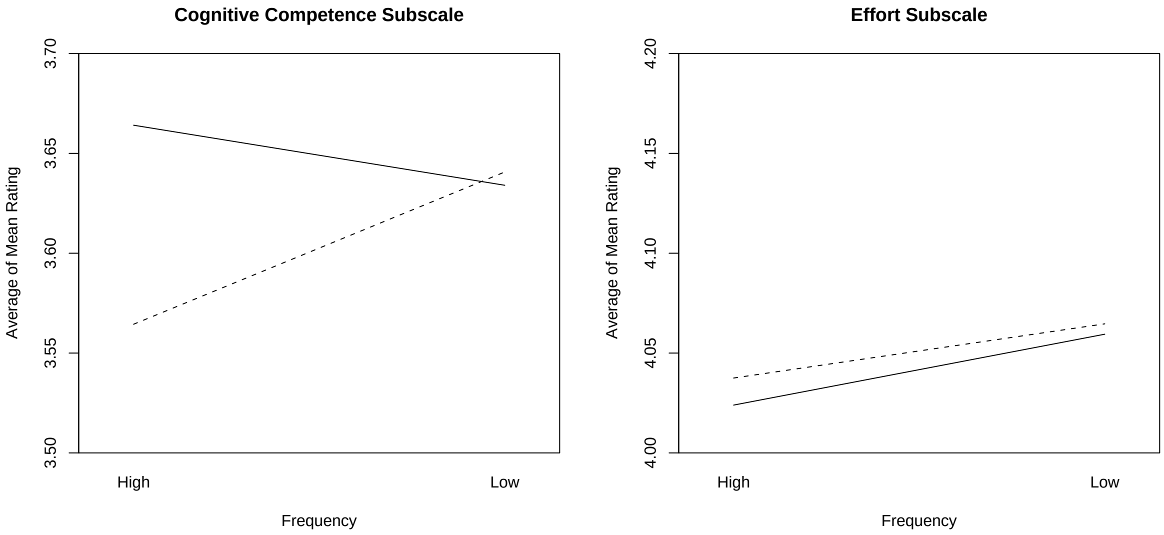

Figure 3.2: Average Mean Post Treatment Ratings by Design Factor for Scales Measuring Cognitive Engagement. In each plot, the solid line corresponds to Clustering *On*, the dashed line to Clustering *Off*. Both plots are scaled to have the same range of 0.2 points on the y-axis.

Hierarchical models were fit using the average ratings for each of these subscales as responses, and following the selection procedure described in the previous subsection. Table 3.10 provides results for the final models. For the model of students' percieved competence in statistical ability, the estimated effects were 0.015 points for Frequency, -0.073 points for Clustering, and 0.027 points for their interaction. For the model of students' effort expended in completing statistical tasks, the estimated effects were nearly zero points for Frequency, -0.054 points for Clustering, and 0.049 points for their interaction. None of the effects for these models were significant. The estimated variance components for each model were small, with the largest being for residual variation: $\hat{\sigma}_{gsi} = 0.056$, $\hat{\sigma}_{lab} \approx 0$, and $\hat{\sigma}_{\epsilon} = 0.556$ for Cognitive Competence and $\hat{\sigma}_{gsi} = 0.070$, $\hat{\sigma}_{lab} = 0.58$, and $\hat{\sigma}_{\epsilon} = 0.687$ for Effort.

Table 3.10: HLM Results for Subscales Measuring Cognitive Engagement

Cognitive Competence Subscale				
	Estimate	Std.Error	DF	P-value
Intercept	3.455	0.065	877	0.000
Pretreatment CAOS	0.010	0.002	877	0.000
Pretreatment Attitudes	0.713	0.047	877	0.000
Grade Point Average: Low	-0.180	0.090	877	0.046
Grade Point Average: High	0.115	0.042	877	0.006
Calculus	0.208	0.048	877	0.000
Lab Start Time: Early Morning	0.155	0.070	23	0.036
Lab Start Time: Late Morning	0.034	0.050	23	0.507
Lab Start Time: Evening	-0.021	0.060	23	0.726
Crossover Sequence 2	-0.025	0.072	17	0.737
Crossover Sequence 3	0.028	0.067	17	0.680
Crossover Sequence 4	-0.009	0.065	17	0.891
Frequency	0.0015	0.022	17	0.749
Clustering	-0.073	0.026	17	0.172
Interaction	-0.027	0.024	17	0.579

Effort Subscale				
	Estimate	Std.Error	DF	P-value
Intercept	4.191	0.080	870	0.000
Pretreatment CAOS	-0.002	0.002	870	0.343
Pretreatment Attitudes	0.252	0.059	870	0.000
Grade Point Average: Low	-0.221	0.111	870	0.047
Grade Point Average: High	0.223	0.052	870	0.000
Year: Freshman	0.086	0.073	870	0.241
Year: Junior	-0.029	0.060	870	0.634
Year: Senior	-0.142	0.073	870	0.052
Gender: Male	-0.278	0.048	870	0.000
Instructor 1	-0.121	0.057	870	0.035
Instructor 3	-0.151	0.125	870	0.227
Instructor 4	-0.145	0.080	870	0.069
Lab Start Time: Early Morning	-0.081	0.095	23	0.402
Lab Start Time: Late Morning	-0.132	0.067	23	0.062
Lab Start Time: Evening	0.004	0.086	23	0.959
Crossover Sequence 2	0.142	0.095	17	0.153
Crossover Sequence 3	0.130	0.088	17	0.157
Crossover Sequence 4	0.026	0.086	17	0.761
Frequency	0.0004	0.029	17	0.994
Clustering	-0.054	0.033	17	0.433
Interaction	0.049	0.031	17	0.437

Note: Estimates reported for Frequency, Clustering, and the Interaction reflect the coding of these factors. That is, since these factors were coded as -1/+1, the estimated regression coefficient was multiplied by two to find the effect of going from the lower level of the factor to the higher level.

Attitudes Towards Clickers

Figure 3.3 plots the mean post treatment ratings by design factor for those statements from the CRLT survey measuring attitude toward clickers. These statements include aspects of both emotional and cognitive engagement, and they are specific to the technology used in this experiment. Again, there there appears to be slight evidence of an interaction between Frequency and Clustering, but of small magnitude.

A hierarchical model confirms that the effects are not significant. The estimated effects were were 0.050 points for Frequency, 0.004 points for Clustering, and 0.069 points for their interaction. Additionally, the estimated variance components are small ($\hat{\sigma}_{gsi} = 0.090$, $\hat{\sigma}_{lab} \approx 0$, and $\hat{\sigma}_{\epsilon} = 0.607$).

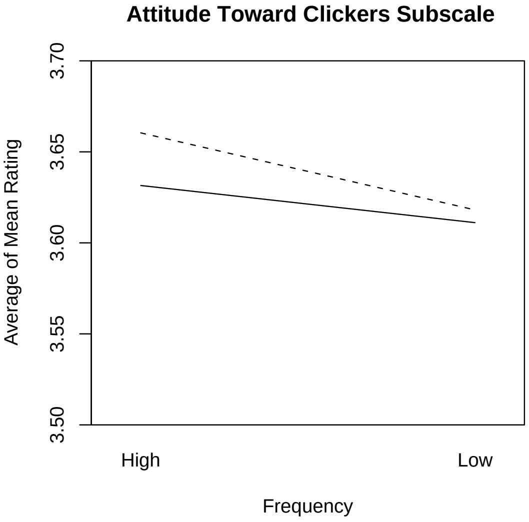

Figure 3.3: Average Mean Post Treatment Ratings by Design Factor for Attitude Toward Clickers. The solid line corresponds to Clustering *On*, the dashed line to Clustering *Off*. The y-axis is scaled to a range of 0.2 points, as in Figures 3.1 and 3.2.

Examining Individual Attitude Statements

In the analyses above, average rating per student was treated as a continuous response variable. While this is common practice, and provides a good idea of "overall" attitudes, it does not account for the fact that the underlying ratings for individual statements are in fact ordinal. To account for this, hierarchical ordinal regressions using the cumulative probit model were run separately for each of the 37 statements on the attitude survey. A modified version of the backward selection procedure presented earlier was used to identify important covariates to include in the model. First, all potential covariates were included model (note that the pretreatment rating for a

Table 3.11: HLM Results for Attitude Toward Clickers Subscale

	Estimate	Std.Error	DF	P-value
Intercept	3.643	0.070	860	0.000
Pretreatment CAOS	0.001	0.002	860	0.604
Pretreatment Attitudes	0.734	0.053	860	0.000
Year: Freshman	0.173	0.061	860	0.005
Year: Junior	0.011	0.053	860	0.833
Year: Senior	-0.028	0.065	860	0.667
Gender: Male	-0.191	0.042	860	0.000
Instructor 1	0.027	0.050	860	0.591
Instructor 3	0.028	0.110	860	0.800
Instructor 4	-0.125	0.070	860	0.073
Crossover Sequence 2	0.054	0.086	17	0.540
Crossover Sequence 3	0.090	0.085	17	0.307
Crossover Sequence 4	0.073	0.082	17	0.390
Frequency	0.050	0.028	17	0.392
Clustering	0.004	0.032	17	0.953
Interaction	0.069	0.030	17	0.258

Note: Estimates reported for Frequency, Clustering, and the Interaction reflect the coding of these factors. That is, since these factors were coded as -1/+1, the estimated regression coefficient was multiplied by two to find the effect of going from the lower level of the factor to the higher level.

particular question was included, rather than average pretreatment attitude rating). Second, any covariates that were not significant at the 10% level were removed from the model, subject to the constraints described earlier. Seven statements showed significant effects of the design factors, using a 5% level for the main effects and a 10% level for the interaction:

1. The clicker questions asked in this lab helped me learn course concepts.
2. I liked using the clickers.
3. I learned more in this lab due to the use of clickers that I would have learned without them.
4. I am scared by statistics.
5. I made a lot of math errors in statistics.
6. I will have no application for statistics in my profession.
7. I use statistics in my everyday life.

For each of these statements, results from the hierarchical ordinal models, as well as the probabilities of giving a particular rating for that statement on the post treatment survey, are provided in Tables 3.5.1 to 3.5.1. In each table, the probabilities are calculated within each treatment group (Team), for a student who earned the average percent correct on the pretreatment CAOS, provided a neutral rating for the corresponding statement on the pretreatment survey, and is in the reference category for each other covariate. For example, in Table 3.5.1, for a female student assigned to the first crossover condition that received the average score on the first CAOS and provided a neutral pretreatment rating to the statement "The clicker questions asked in this lab helped me learn course concepts," the probability of providing a post treatment rating of "Agree" to the same statement is 0.50 if that student was assigned to the Green Team, 0.46 if she was assigned to the Blue Team, etc.

Table 3.12: Results for the Statement: *The clicker questions asked in this lab helped me learn course concepts*

	Estimate	Std.Error	P-value
Threshold1	0.624	0.289	0.031
Threshold2	1.487	0.281	0.000
Threshold3	2.172	0.283	0.000
Threshold4	4.050	0.300	0.000
Pretreatment CAOS	0.008	0.003	0.012
Pretreatment Rating	0.630	0.051	0.000
Gender: Male	-0.230	0.079	0.003
Crossover Sequence 2	0.350	0.136	0.010
Crossover Sequence 3	0.210	0.133	0.113
Crossover Sequence 4	0.118	0.128	0.355
Frequency	0.109	0.045	0.015
Clustering	-0.004	0.050	0.935
Interaction	0.065	0.047	0.165

	Probability[b] of Post treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.05	0.17	0.24	0.50	0.04
Blue	0.06	0.19	0.26	0.46	0.03
Orange	0.04	0.15	0.23	0.53	0.05
Yellow	0.03	0.13	0.22	0.56	0.06

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

As can be seen in the top half of these tables, there is variation both in the included covariates and the significant treatment factor, perhaps reflecting differences in the statements. Some consistencies can be seen, however. For example, when the effect of Frequency is significant at the 5% level, it is positive, indicating that

Table 3.13: Results for the Statement: *I liked using clickers*

	Estimate	Std.Error	P-value
Threshold1	1.192	0.238	0.000
Threshold2	2.062	0.239	0.000
Threshold3	3.103	0.247	0.000
Threshold4	4.709	0.265	0.000
Pretreatment CAOS	0.008	0.003	0.007
Pretreatment Rating	0.842	0.043	0.000
Year: Freshman	0.137	0.107	0.202
Year: Junior	0.015	0.093	0.875
Year: Senior	-0.249	0.113	0.027
Gender: Male	-0.182	0.077	0.018
Crossover Sequence 2	0.136	0.118	0.251
Crossover Sequence 3	0.010	0.114	0.932
Crossover Sequence 4	0.130	0.111	0.240
Frequency	0.058	0.039	0.130
Clustering	0.096	0.044	0.028
Interaction	-0.020	0.040	0.628

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.05	0.18	0.39	0.35	0.03
Blue	0.03	0.13	0.36	0.42	0.05
Orange	0.04	0.14	0.37	0.41	0.04
Yellow	0.03	0.12	0.35	0.45	0.05

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

asking more clicker questions is better. When the interaction between Frequency and Clustering is significant at the 10% level, it is negative (see Tables 3.5.1 and 3.5.1), consistent with Hypothesis 1 in Section 3.2. Interestingly, the effect of Clustering (when significant at the 5% level) is positive for a statement pertaining to clickers ("I liked using the clickers") and negative for a statement pertaining to statistics ("I made a lot of math errors in statistics"). It would seem unlikely that asking all clicker questions in a row would increase the number math errors made by a student; of course, it is plausible that this relationship is simply spurious. Pertaining to clickers specifically, it seems as though students liked using them more when the clicker questions were well-integrated into the lesson rather than asked in a row. Considering the probabilities provided in the second half of these tables, the largest probability of moving from a pretreatment rating of "Neutral" to a post treatment rating of "Agree" is generally largest for the Yellow Team (Frequency =

Table 3.14: Results for the Statement: *I learned more using clickers than I would have learned without them*

	Estimate	Std.Error	P-value
Threshold1	0.565	0.240	0.018
Threshold2	1.611	0.239	0.000
Threshold3	2.604	0.246	0.000
Threshold4	4.017	0.259	0.000
Pretreatment CAOS	0.008	0.003	0.010
Pretreatment Rating	0.631	0.044	0.000
Crossover Sequence 2	0.139	0.114	0.223
Crossover Sequence 3	0.033	0.111	0.765
Crossover Sequence 4	0.074	0.107	0.489
Frequency	0.116	0.037	0.002
Clustering	-0.027	0.042	0.515
Interaction	0.054	0.039	0.170

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.05	0.21	0.38	0.33	0.04
Blue	0.06	0.25	0.38	0.28	0.03
Orange	0.03	0.19	0.37	0.36	0.05
Yellow	0.03	0.17	0.36	0.38	0.06

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

High, Clustering = *Off*). For all teams and across all statements, the probability of making this improved rating is encouragingly high—ranging from 28% to 56% and often higher than making the change to a negative rating of “Strongly Disagree” or “Disagree.” An exception to this (not too surprisingly) is the last statement, “I use statistics in my everyday life.” Also calculated, but not shown here, were the probabilities of improving from a pretreatment rating of “Disagree” to a post treatment rating of “Agree”. While these probabilities were understandably lower than those presented in the tables, the were still encouraging—ranging from 12% to 36% across all teams and statements (excluding the last statement).

3.5.2 Behavioral engagement outcome: Clicker use and External Incentive

GSIs were randomly assigned to one of four treatment sequences based on possible combinations of the three levels of External Incentive the constraint that a switch between required (External Incentive = *High*) and optional (External Incentive = *Moderate* or *Low*) clicker use be made only once during the semester. The resulting

Table 3.15: Results for the Statement: *I am scared by statistics*

	Estimate	Std.Error	P-value
Threshold1	0.856	0.216	0.000
Threshold2	2.096	0.216	0.000
Threshold3	2.826	0.223	0.000
Threshold4	4.355	0.240	0.000
Pretreatment CAOS	0.014	0.003	0.000
Pretreatment Rating	0.674	0.039	0.000
Grade Point Average: Low	-0.358	0.182	0.049
Grade Point Average: High	0.092	0.085	0.280
Calculus	0.270	0.097	0.005
Crossover Sequence 2	-0.036	0.118	0.757
Crossover Sequence 3	0.051	0.114	0.654
Crossover Sequence 4	-0.237	0.111	0.033
Frequency	-0.005	0.038	0.897
Clustering	-0.073	0.044	0.093
Interaction	-0.096	0.041	0.018

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.03	0.24	0.28	0.41	0.05
Blue	0.03	0.22	0.27	0.42	0.06
Orange	0.02	0.19	0.26	0.46	0.07
Yellow	0.04	0.28	0.28	0.36	0.04

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

sequences were:

1. Low – Moderate – High
2. Moderate – Low – High
3. High – Low – Moderate
4. High – Moderate – Low

Two analyses of clicker use were performed. For the first, clicker use was defined the number of students answering at least one clicker question during a given week, weighted to account for varying lab sizes. For the second, clicker use was defined as the number of students answering at least 50% of the clicker questions during a given week, again weighted to account for varying lab sizes. Each of these is discussed in turn below.

Table 3.16: Results for the Statement: *I made a lot of math errors in statistics*

	Estimate	Std.Error	P-value
Threshold1	0.038	0.234	0.870
Threshold2	1.400	0.230	0.000
Threshold3	2.080	0.235	0.000
Threshold4	3.633	0.247	0.000
Pretreatment CAOS	0.010	0.003	0.002
Pretreatment Rating	0.478	0.042	0.000
Grade Point Average: Low	-0.320	0.181	0.077
Grade Point Average: High	0.145	0.083	0.082
Year: Freshman	0.122	0.105	0.244
Year: Junior	-0.157	0.093	0.089
Year: Senior	0.076	0.112	0.498
Calculus	0.181	0.096	0.059
Crossover Sequence 2	-0.107	0.116	0.358
Crossover Sequence 3	0.057	0.113	0.612
Crossover Sequence 4	-0.212	0.110	0.055
Frequency	0.011	0.038	0.769
Clustering	-0.097	0.043	0.025
Interaction	-0.040	0.040	0.315

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.02	0.25	0.26	0.42	0.05
Blue	0.03	0.28	0.26	0.38	0.04
Orange	0.02	0.22	0.25	0.45	0.06
Yellow	0.04	0.30	0.26	0.37	0.04

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

Students answering at least one clicker question

Table 3.5.2 shows the proportion of students in each sequence (as defined above) who answered at least one clicker question for a particular week of the semester (recall that weeks 4–12 of the experimental semester defined the treatment period). Figure 3.4 plots the same information. All sequences show some decrease in the proportion of users over the course of the treatment period, though the magnitude of these decreases is often small. Sequence 4, where the level of External Incentive steadily decreases from the beginning to the end of the treatment period, shows the largest decline in clicker use—from over 90% use when clickers were required (External Incentive = *High*) to less than 70% use when clicker use was anonymous (External Incentive = *Low*). This trend certainly is not surprising, as grades are a powerful motivator for students to use clickers when they were required, but there is

Table 3.17: Results for the Statement: *I will have no application for statistics in my profession*

	Estimate	Std.Error	P-value
Threshold1	1.013	0.264	0.000
Threshold2	2.033	0.259	0.000
Threshold3	2.983	0.265	0.000
Threshold4	4.691	0.282	0.000
Pretreatment CAOS	0.011	0.003	0.001
Pretreatment Rating	0.732	0.049	0.000
Year: Freshman	-0.066	0.107	0.539
Year: Junior	-0.029	0.095	0.758
Year: Senior	0.239	0.115	0.039
Gender: Male	-0.316	0.078	0.000
Calculus	0.261	0.097	0.007
Crossover Sequence 2	0.259	0.120	0.031
Crossover Sequence 3	0.199	0.116	0.085
Crossover Sequence 4	0.024	0.112	0.833
Frequency	-0.025	0.039	0.514
Clustering	-0.009	0.044	0.841
Interaction	-0.069	0.041	0.089

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.04	0.20	0.36	0.38	0.03
Blue	0.03	0.17	0.35	0.42	0.03
Orange	0.04	0.18	0.35	0.40	0.03
Yellow	0.05	0.21	0.36	0.35	0.02

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

no accountability when clicker use was anonymous. For sequence 3, there is a sharp drop-off in clicker use during the final period of the crossover experiment (week 10–12), when the level of External Incentive was *Moderate* for this group. Apparently, even the tracking of individual student's clicker use was not enough incentive to use clickers by the end of the term.

To explore if External Incentive had any significant effect on behavioral engagement, as measured through students' self-selected clicker use, a hierarchical linear model with nested random effects for GSI and lab section was fit accounting for sequence, period and week effects. It was not possible to include individual student-level covariates in this model, as there was no way to identify individual students under the *Low* level of External Incentive. Therefore, the model fitting procedure for this outcome was:

1. The GSI and lab level covariates described in Section 3.4 were initially included

Table 3.18: Results for the Statement: *I use statistics in my everyday life*

	Estimate	Std.Error	P-value
Threshold1	0.274	0.233	0.240
Threshold2	1.811	0.233	0.000
Threshold3	2.930	0.240	0.000
Threshold4	4.477	0.263	0.000
Pretreatment CAOS	0.002	0.003	0.433
Pretreatment Rating	0.602	0.045	0.000
Calculus	0.338	0.095	0.000
Lab Start Time: Early Morning	-0.013	0.125	0.915
Lab Start Time: Late Morning	0.011	0.093	0.909
Lab Start Time: Evening	-0.214	0.109	0.050
Crossover Sequence 2	0.138	0.124	0.264
Crossover Sequence 3	0.060	0.113	0.599
Crossover Sequence 4	0.033	0.110	0.760
Frequency	0.077	0.038	0.042
Clustering	0.052	0.043	0.233
Interaction	0.024	0.040	0.554

	Probability[b] of Post Treatment Rating				
Team[a]	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Green	0.06	0.43	0.37	0.13	0.00
Blue	0.05	0.42	0.38	0.14	0.00
Orange	0.05	0.40	0.39	0.15	0.01
Yellow	0.04	0.36	0.41	0.19	0.01

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

[b] Probabilities are calculated for students with the average score of the pretreatment CAOS, a neutral rating for the corresponding pretreatment statement, and in the reference category for all other covariates.

Table 3.19: Proportion of Students Answering At Least One Clicker Question

	Week								
Sequence[a]	4	5	6	7	8	9	10	11	12
1	0.916	0.911	0.869	0.850	0.835	0.818	0.827	0.869	0.869
2	0.896	0.921	0.921	0.869	0.849	0.869	0.865	0.883	0.876
3	0.918	0.938	0.913	0.851	0.831	0.817	0.826	0.717	0.722
4	0.911	0.953	0.937	0.824	0.849	0.849	0.741	0.689	0.640

[a] The sequences are: 1 (Low-Mod-High External Incentive); 2 (Mod-Low-High); 3 (High-Low-Mod); 4 (High-Mod-Low)

in the model. Covariates that were insignificant at the 10% level were individually dropped from the model until only significant covariates remained.

- Indicators of the treatment group from the factorial experiment that a particular GSI had been randomized to could not be dropped. These were included in the model to account for any effects of the design factors Frequency and Clustering, which were not of particular interest when estimating the effects of External Incentive but needed to be accounted for.

- Indicators of the *Moderate* and *High* levels of External Incentive (using the *Low* level as the reference group) could not be dropped.

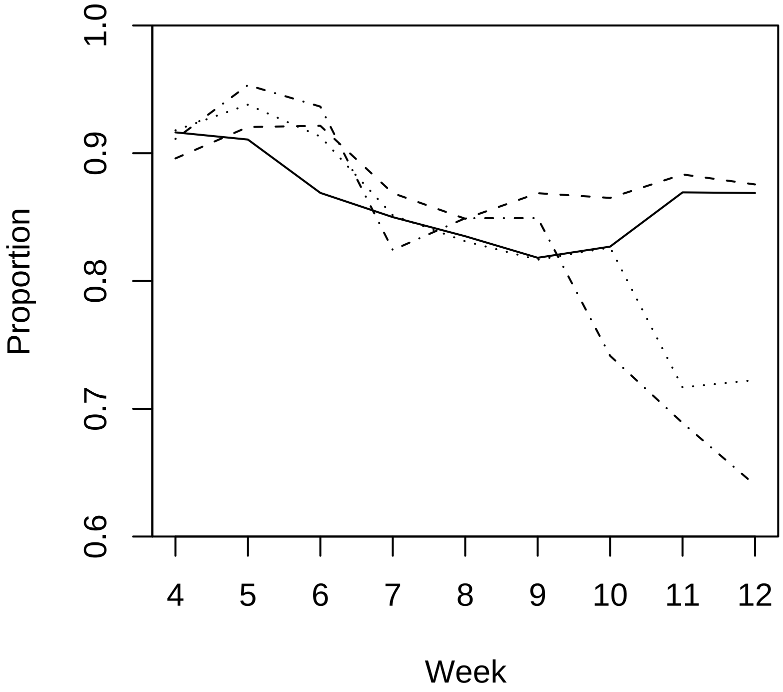

Figure 3.4: Proportion of Students Answering At Least One Clicker Question. The solid represents the proportion, for each week of the treatment period, of students in sequence 1 (Low-Mod-High External Incentive) who answered at least one clicker question; the dashed line represents the corresponding proportions for students in sequence 2 (Mod-Low-High); the dotted line represents sequence 3 (High-Low-Mod); and the dashed and dotted line represents sequence 4 (High-Mod-Low).

2. After all non-significant covariates were removed, the Akaike information criterion (AIC) for the reduced model was compared to the AIC for the full model. The model with the smaller AIC was taken as the final model.

The response for this model was the number of students in each lab section answering at least one clicker question for a given week, with weights equal to the number of students in attendance for that section that week. (When attendance numbers were missing for a particular lab section on a given week, weights were set equal to the number of students enrolled in that section after the drop/add deadline. Since this number should be greater than or equal to actual attendance figures each week, this should be a conservative estimate of the appropriate sample size.) Results from this model are shown in Table 3.20. It can be seen that the estimated number of clicker users significantly increases with each level of External Incentive: 0.751 and 1.792 additional students used clickers to answer at least one question under the *Moderate* and *High* levels, respectively, of External Incentive as compared to under the *Low* level. The largest sources of variation is due to GSI, with variation due to lab a close second: $\hat{\sigma}_{gsi} = 1.958$, $\hat{\sigma}_{lab} = 1.779$ and $\hat{\sigma}_{\epsilon} = 0.478$.

Table 3.20: HLM Results for Behavioral Engagement—Number of Students Answering At Least One Clicker Question

	Estimate	Std.Error	DF	P-value
Intercept	20.91571	1.497912	308	0.000
Team: Blue	-0.718	1.513	17	0.641
Team: Yellow	0.027	1.493	17	0.986
Team: Orange	0.857	1.409	17	0.551
Crossover Sequence 2	0.614	1.537	17	0.694
Crossover Sequence 3	-1.463	1.522	17	0.350
Crossover Sequence 4	-1.465	1.451	17	0.327
Period 2	-1.069	0.571	308	0.062
Period 3	-3.588	0.940	308	0.000
Week	0.032	0.150	308	0.831
Incentive: Moderate	0.751	0.299	308	0.013
Incentive: High	1.792	0.332	308	0.000

Students answering at least 50% of the clicker questions

Table 3.5.2 and Figure 3.5 show the proportion of students in each sequence who answered at least 50% of the clicker questions for a particular week of the semester. Due to the stricter definition for clicker use, the proportions are understandably lower for each sequence in each week than with the previous definition of clicker use. As previously, however, all sequences show some decrease in the proportion of users over the course of the treatment period, with Sequences 3 and 4 showing the largest declines. Here, the sharp decline in clicker use for sequence 3 occurs during the second period of the crossover experiment (weeks 7–9, when the level of External Incentive was *Low* for this group) and continues through the final period (weeks 10–12; *Moderate* External Incentive).

Table 3.21: Proportion of Students Answering At Least 50% of the Clicker Questions

	Week								
Sequence[a]	4	5	6	7	8	9	10	11	12
1	0.865	0.857	0.841	0.825	0.804	0.791	0.808	0.853	0.844
2	0.879	0.900	0.901	0.838	0.824	0.838	0.834	0.862	0.853
3	0.843	0.899	0.905	0.728	0.760	0.747	0.768	0.674	0.684
4	0.893	0.946	0.921	0.816	0.806	0.801	0.683	0.644	0.597

[a] The sequences are: 1 (Low-Mod-High External Incentive); 2 (Mod-Low-High); 3 (High-Low-Mod); 4 (High-Mod-Low)

Again a hierarchical linear model, weighted by the number of students in attendance for a particular lab section and week, was fit. Here the response was the number of students in each lab section answering at least 50% of the clicker questions for a given week. Results from this model are shown in Table 3.22. Consistent with previous findings, the estimated number of clicker users significantly increases with each level of External Incentive after accounting for sequence, period, and week effects: 1.275 and 2.347 additional students used clickers to answer at least 50% of the clicker questions under the *Moderate* and *High* levels, respectively, of External Incentive as compared to under the *Low* level. Here GSI and lab represent almost equally large sources of variation: $\hat{\sigma}_{gsi} = 1.991$, $\hat{\sigma}_{lab} = 1.941$ and $\hat{\sigma}_{\epsilon} = 0.561$.

Table 3.22: HLM Results for Behavioral Engagement—Number Answering At Least 50% of the Clicker Questions

	Estimate	Std.Error	DF	P-value
Intercept	19.407	1.626	308	0.000
Team: Blue	-0.827	1.587	17	0.609
Team: Yellow	-0.968	1.566	17	0.545
Team: Orange	0.619	1.477	17	0.680
Crossover Sequence 2	0.658	1.615	17	0.689
Crossover Sequence 3	-1.900	1.598	17	0.251
Crossover Sequence 4	-1.712	1.521	17	0.276
Period 2	-1.780	0.671	308	0.008
Period 3	-4.744	1.105	308	0.000
Week	0.206	0.177	308	0.245
Incentive: Moderate	1.275	0.351	308	0.000
Incentive: High	2.347	0.390	308	0.000

3.5.3 Learning outcome: The Comprehensive Assessment of Outcomes in a first course in Statistics

The primary measure of learning for this experiment was the Comprehensive Assessment of Outcomes in a first course in Statistics (CAOS) instrument. Students

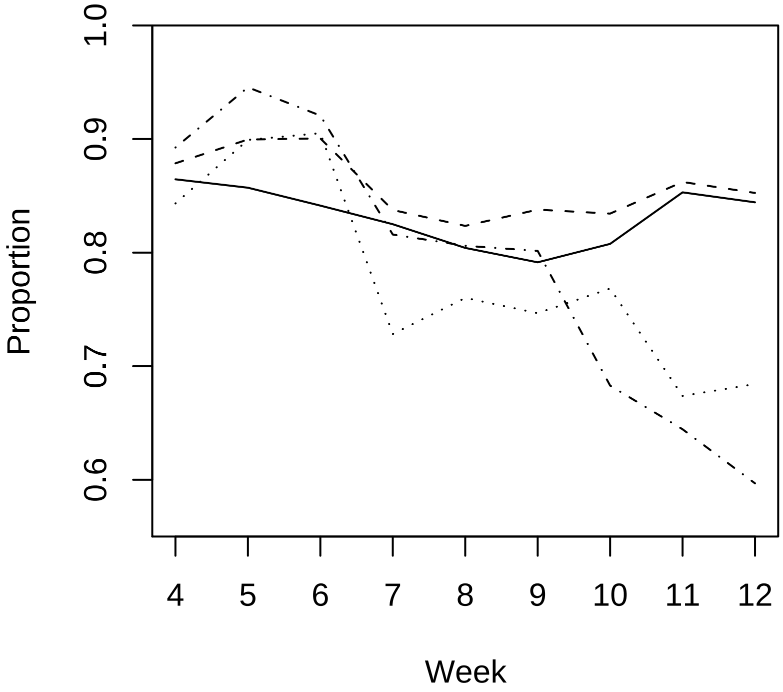

Figure 3.5: Proportion of Students Answering At Least 50% of the Clicker Questions. The solid represents the proportion, for each week of the treatment period, of students in sequence 1 (Low-Mod-High External Incentive) who answered at least 50% of the clicker questions; the dashed line represents the corresponding proportions for students in sequence 2 (Mod-Low-High); the dotted line represents sequence 3 (High-Low-Mod); and the dashed and dotted line represents sequence 4 (High-Mod-Low).

completed CAOS four times throughout the term. The first, which was considered as a pretreatment measure of statistical understanding, took place during the third lab session. This was done to accommodate the drop/add period at the start of the semester, during which the course roster changes often. After the drop/add deadline had passed, course enrollment was fixed (with the exception of a handful of students who dropped late), making it more feasible to regularly collect measurements on the students. By the time they completed the first CAOS, students had learned about graphical and numeric data summaries, including the mean, standard deviation, quartiles, range, histograms and boxplots. Based on this, students could have correctly answered about 30% of the 40 CAOS questions; in actuality, students on average correctly answered about 52% of the questions at this time (see Table 3.23). All students were required to complete the first and final administrations of CAOS. Completion of the second and third installments was optional; students were awarded a small amount of extra credit for answering most of the questions. Extra credit was added to the corresponding midterm exam score (i.e. two points were added to the first midterm score for completing the second CAOS; two points were added to the second midterm score for completing the third CAOS). Descriptive statistics for each of the CAOS exams, for the entire sample (Overall) and by treatment group (Team), are given in Table 3.23. While the values of Cronbach's α are just below the conventional threshold of 0.70 for the pretreatment CAOS, the values improve to acceptable levels for the remaining time points. The treatment groups had roughly equivalent scores on the first CAOS, with the Green Team (Frequency = *Low*, Clustering = *Off*) having a slightly higher mean than the other teams. Overall, the average CAOS score increased at each assessment period, increasing by 13.7% (equivalent to 5 and a half points) from pre- to post treatment. It can also be seen that the number of students completing the second and third CAOS assessments was quite a bit lower than the number completing the first and final CAOS. For this reason, the final installment of CAOS was the primary outcome of interest, adjusting for the pretreatment CAOS score.

Figure 3.6 plots the average percent correct on the final CAOS by treatment factor. Interestingly, the lines in this picture appear parallel, indicating that there is no interaction between Frequency and Clustering. The *Low* level of Frequency always appears to be better than the *High* level, and the *Off* level of clustering always appears to be better than *On*. To test if Frequency and Clustering had

Table 3.23: Descriptive Statistics for CAOS

	Team[a]	Cronbach's α	Min	Median	Mean (SD)	Max	N
First CAOS	Overall	0.67	7.5	50.0	52.1 (12.3)	92.5	1163
	Green	0.69	17.5	55.0	54.0 (12.6)	87.5	1150
	Blue	0.67	7.5	50.0	51.5 (12.4)	92.5	1153
	Orange	0.69	25.0	50.0	51.7 (12.6)	85.0	1158
	Yellow	0.62	20.0	50.0	51.0 (11.5)	85.0	1157
Second CAOS	Overall	0.79	2.5	60.0	58.7 (14.9)	92.5	758
	Green	0.77	22.5	60.0	59.4 (14.3)	90.0	645
	Blue	0.80	22.5	60.0	58.3 (15.5)	92.5	657
	Orange	0.79	2.5	60.0	59.3 (14.9)	87.5	650
	Yellow	0.78	5.0	57.5	57.9 (14.9)	87.5	641
Third CAOS	Overall	0.77	20.0	62.5	61.2 (14.1)	90.0	688
	Green	0.76	27.5	65.0	62.2 (13.6)	90.0	574
	Blue	0.78	20.0	62.5	62.5 (14.0)	90.0	568
	Orange	0.79	20.0	60.0	60.2 (14.7)	90.0	554
	Yellow	0.76	30.0	60.0	60.0 (14.2)	90.0	547
Fourth CAOS	Overall	0.76	20.0	67.5	65.7 (13.1)	97.5	1128
	Green	0.75	27.5	67.5	66.4 (12.8)	95.0	1112
	Blue	0.74	20.0	67.5	67.3 (12.6)	95.0	1118
	Orange	0.79	25.0	65.0	64.2 (14.1)	92.5	1101
	Yellow	0.73	20.0	65.0	65.2 (12.6)	97.5	1112

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

significant effects, a hierarchical linear model was fit including nested random effects for GSI and lab. To identify important confounding variables the same backward selection procedure presented in Section 3.5.1 was used, with the exception that the pretreatment measure of attitudes toward statistics and clickers included was the average rating from the entire attitude survey rather than ratings for a particular subset of questions.

Table 3.24 provides the final model produced by this fitting procedure. The response for this model is the percent correct on the final CAOS. After adjusting for several important confounders, the main effect of Frequency is estimated to be -1.370 percent; the main effect of Clustering is estimated to be 1.605 percent; and the effect of the interaction is estimated to be -1.494 percent. These estimated effects all correspond to a change of less than 1 point (out of 40 points possible) on the final CAOS. The interaction is significant at the 10% level; the main effects of Frequency and Clustering are both insignificant at the 5% level. This analysis indicates that, holding all else equal, asking a low number (3–4) of clicker questions

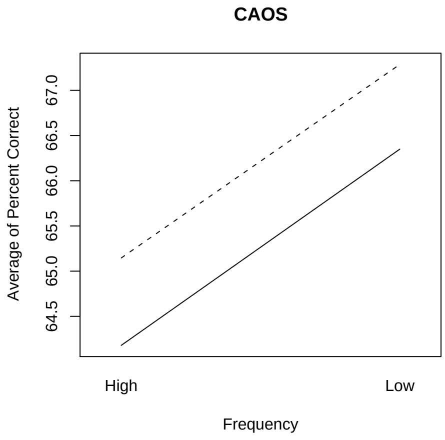

Figure 3.6: Average Percent Correct for Final CAOS by Treatment Group. The solid line corresponds to Clustering *On*, the dashed line to Clustering *Off*.

and incorporating those questions throughout a class led to an increase of 4.469 percent correct, or roughly 2 points, on the final CAOS (as compared to asking a high number [more than 6] of clicker questions and asking those questions consecutively [i.e. in a "cluster"]).

To ensure that the model fitting process did not produce a model that was too sample-specific, a simple validation procedure was used. Specifically, the sample of complete cases was divided into quarters, and a different quarter was excluded from each of four subsamples of data. The covariate selection procedure was repeated using each of the resulting three-quarter subsamples and the final validation models produced were examined for consistency with the final model presented in Table 3.24. The overall substantive conclusions about the magnitude and significance of the design factors was consistent for each of these four validation models (not shown).

Question-level Analysis of CAOS

Since the 40 CAOS questions were not of equal difficulty, several descriptive analyses were undertaken to explore the performance of the treatment groups (Team)

Table 3.24: HLM Results for Percent Correct on Final CAOS

	Estimate	Std.Error	DF	P-value
Intercept	64.280	1.091	876	0.000
Pretreatment CAOS	0.599	0.029	876	0.000
Pretreatment Attitudes	1.773	0.832	876	0.034
Grade Point Average: Low	-3.679	1.562	876	0.019
Grade Point Average: High	2.752	0.730	876	0.000
Year: Freshman	-2.699	1.028	876	0.009
Year: Junior	-2.260	0.848	876	0.008
Year: Senior	1.224	1.026	876	0.233
Gender: Male	1.575	0.671	876	0.019
Instructor 1	1.797	0.805	876	0.026
Instructor 3	-0.114	1.761	876	0.948
Instructor 4	0.749	1.123	876	0.505
Lab Start Time: Early Morning	2.596	1.305	23	0.059
Lab Start Time: Late Morning	2.516	0.934	23	0.013
Lab Start Time: Evening	0.698	1.187	23	0.562
Crossover Sequence 2	-0.940	1.275	17	0.471
Crossover Sequence 3	0.563	1.177	17	0.638
Crossover Sequence 4	-0.913	1.146	17	0.437
Frequency	-1.370	0.395	17	0.101
Clustering	1.605	0.448	17	0.091
Interaction	-1.494	0.413	17	0.088

Note: Estimates reported for Frequency, Clustering, and the Interaction reflect the coding of these factors. That is, since these factors were coded as -1/+1, the estimated regression coefficient was multiplied by two to find the effect of going from the lower level of the factor to the higher level.

by question. Figure 3.7 shows the proportion of correct responses to each of the 40 questions. In the plot there are four points for each question, one for each team. Regressions lines provide an idea of the average performance for each team. The line that stands out the most is the solid line, which corresponds to the Blue Team (Frequency=*Low*,Clustering=*Off*), indicating that asking a few clicker questions throughout a class results in the highest percentage of correct responses, on average. To look at the team performances on each question in more detail, questions were grouped based on topic. The resulting topics were:

- **sampDist**: Sampling distribution (Questions 16,17,32,34,35)
- **pvalue**: Interpretation of p-value (19,25-27)
- **confInt**: Confidence intervals (28-31)
- **data**: Making sense of data (11-13,18)
- **reg**: Regression - dangers of extrapolation (39)

- **dist**: Understanding distribution (1,3-5)
- **hist**: Reading a histogram (6,33)
- **boxplot**: Reading a boxplot (2,9-10)
- **gatherData**: Gathering data (7,38,22,24)
- **stdDev**: Understanding standard deviation (8,14,15)
- **cor**: Correlation (20,21)
- **practSig**: Practical significance (23)
- **chiSq**: Chi-square / categorical comparisons (36)
- **permTest**: Permutation test / simulation (37)
- **hypTest**: Hypothesis test conclusion (40)

For each student, the proportion of questions answered correctly for each topic was calculated. These proportions were then averaged over all students in a team to get the team proportion correct. Figure 3.8 plots these values by topic. For each topic, the teams show improvement from pre- to post- treatment. (The only exception to this was on the question involving simulation and permutation tests [permTest], which was not explicitly covered during the course.) While the team proportions do not differ substantially, there are a few topics for which the Blue Team performed best. These topics included: confidence intervals, making sense of data, understanding distribution, reading a histogram, and gathering data.

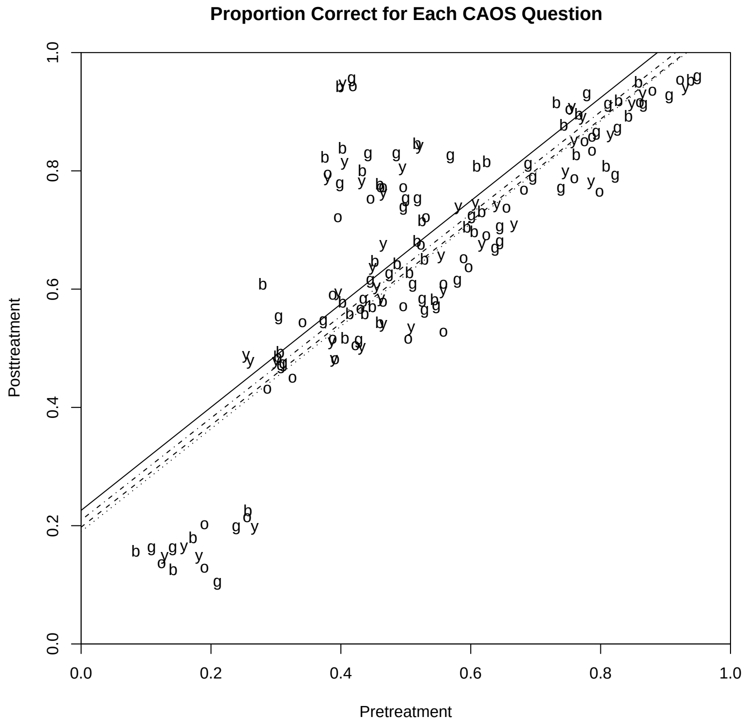

Figure 3.7: Proportion of Correct Responses for Each CAOS Question by Team. Plotting character corresponds to team name: g=Green (Frequency=*Low*, Clustering=*On*); b=Blue (*Low*, *Off*); o=Orange (*High*, *On*); y=Yellow (*High*, *Off*). Linear regression lines provide an idea of the average performance for each group. The solid line corresponds to the Blue Team. The dotted & dashed line corresponds to the Yellow Team. The lines corresponding to the Green Team (dashed) and the Orange Team (dotted) are nearly indistinguishable.

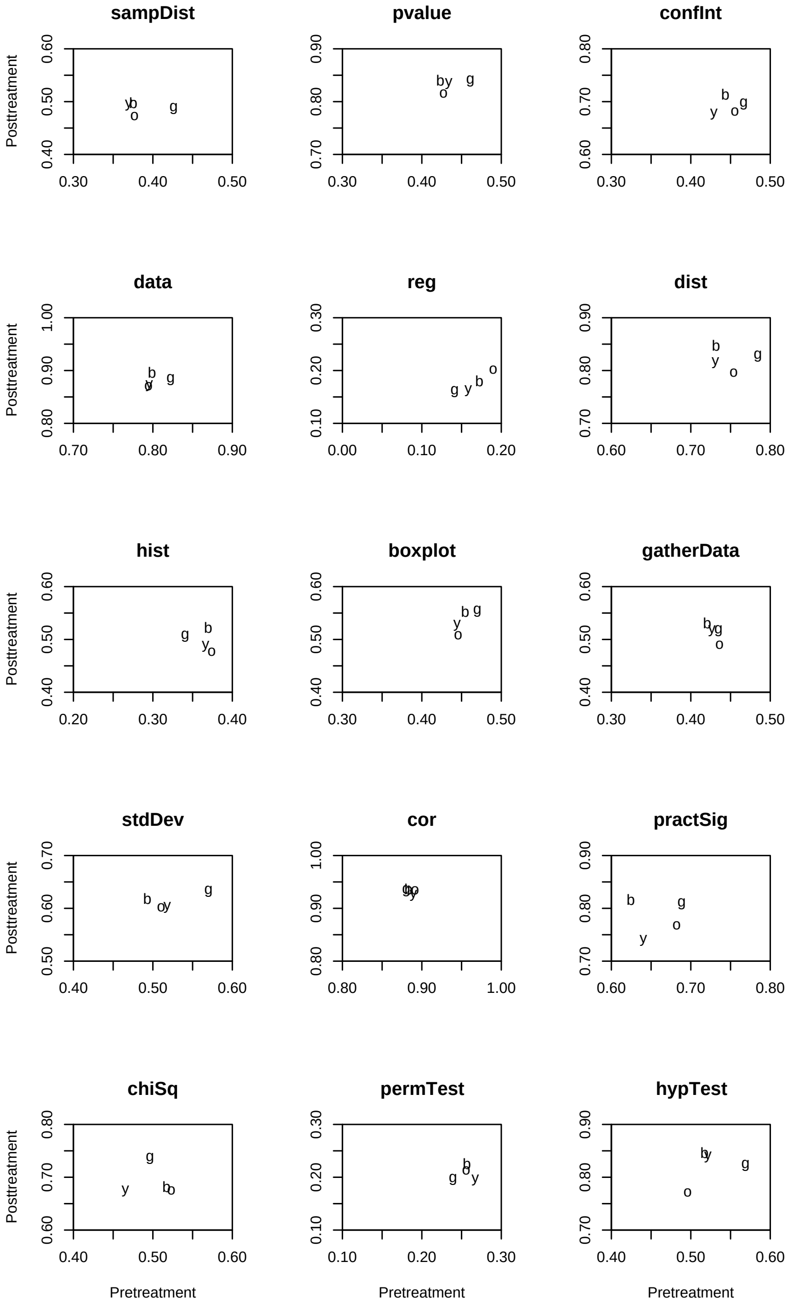

Figure 3.8: Proportion of Correct Responses for Each CAOS Topic by Team. Plotting character corresponds to team name: g=Green (Frequency=*Low*, Clustering=*On*); b=Blue (*Low*, *Off*); o=Orange (*High*, *On*); y=Yellow (*High*, *Off*). Each axis has been scaled to have the same range of 0.20, or 20%.

3.5.4 Learning outcome: ARTIST topic scales

Table 3.25 provides descriptive statistics for each of the four ARTIST topic scales—Normal Distribution, Sampling Distributions, Confidence Intervals, and Significance Tests—for the entire sample (Overall) and by treatment group (Team). The values of Cronbach's α for each scale are notably low—only the scores for the Sampling Distribution scale even approach the acceptable threshold of 0.70. Such low reliabilities might indicate that students did not take these assessments very seriously, or try very hard when answering the questions. Each topic scale was administered at the beginning of a lab session, with students getting between 10 and 15 minutes to answer all questions. They were graded informally—students received a portion of the day's participation points for completing the scale online. Interestingly, though, students performed very well on these scales—the mean and the median scores were well above the 60% mark for each. The online order of the questions and answer choices were not randomized; it is possible then that, given their low-stakes nature, students tended to "work together" more than they should have. While the overall scores were very good, it should be noted that, the Blue team (Frequency = *Low*, Clustering = *Off*) had the highest average score for each topic scale.

Figure 3.9 shows the average percent correct for each of the topic scales by treatment factor. Several plots show evidence of an interaction. In nearly every case, the the *Off* level of Clustering appears to be better than *On*, and the magnitude of this difference is often larger when Frequency is at the *Low* level. However, for each scale, hierarchical models using percent correct as the response did not show significant effects for Frequency, Clustering, or their interaction (results not shown). Given the extremely low reliability shown in Table 3.25, this is not surprising. Because of this, further analysis of the topic scale data was not conducted.

Table 3.25: Descriptive Statistics for the ARTIST Topic Scales

	Team[a]	Cronbach's α	Min	Median	Mean (SD)	Max	N
Normal Distribution (15 Questions)	Overall	0.47	0.0	62.5	64.4 (20.4)	100	1109
	Green	0.45	12.5	62.5	65.0 (20.1)	100	1089
	Blue	0.43	12.5	62.5	66.2 (19.1)	100	1083
	Orange	0.51	12.5	62.5	63.9 (20.7)	100	1089
	Yellow	0.49	0.0	62.5	62.8 (21.3)	100	1087
Sampling Distribution (15 Questions)	Overall	0.64	13.3	66.7	65.7 (18.1)	100	1070
	Green	0.67	13.3	66.7	64.9 (18.8)	100	1050
	Blue	0.67	13.3	66.7	67.2 (18.5)	100	1048
	Orange	0.61	13.3	66.7	65.3 (17.5)	100	1009
	Yellow	0.61	20.0	66.7	65.4 (17.6)	100	1046
Confidence Intervals (10 Questions)	Overall	0.54	10.0	70.0	70.4 (19.1)	100	1098
	Green	0.52	10.0	70.0	70.9 (18.6)	100	1075
	Blue	0.50	10.0	70.0	72.2 (18.3)	100	1074
	Orange	0.55	10.0	70.0	68.6 (19.6)	100	1067
	Yellow	0.56	10.0	70.0	69.8 (19.6)	100	1077
Significance Tests (10 Questions)	Overall	0.50	0.0	70.0	66.4 (19.1)	100	1076
	Green	0.52	0.0	70.0	66.2 (19.8)	100	1041
	Blue	0.47	20.0	70.0	68.2 (18.2)	100	1054
	Orange	0.48	10.0	70.0	65.3 (18.6)	100	1034
	Yellow	0.53	10.0	70.0	66.0 (19.7)	100	1054

[a] The teams are: Green (Frequency=*Low*, Clustering=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

3.6 Discussion: The Effect of Clickers on Engagement and Learning

In Section 3.2, the relevant research questions and hypotheses for the experiment were presented as:

RQ1. Can you "overdose" on clickers by asking too many questions?

RQ2. What is the best way to distribute clicker questions throughout a class session?

H1. There will be a negative effect of clicker overuse—too many clicker questions asked consecutively.

RQ3. Are students motivated to use clickers even when it is neither required nor monitored?

Discussion about each follows.

3.6.1 Discussion of RQ1.

Table 3.6.1 shows the estimated main effects and standard errors for Frequency, Clustering, and their interaction from the hierarchical analyses of engagement and

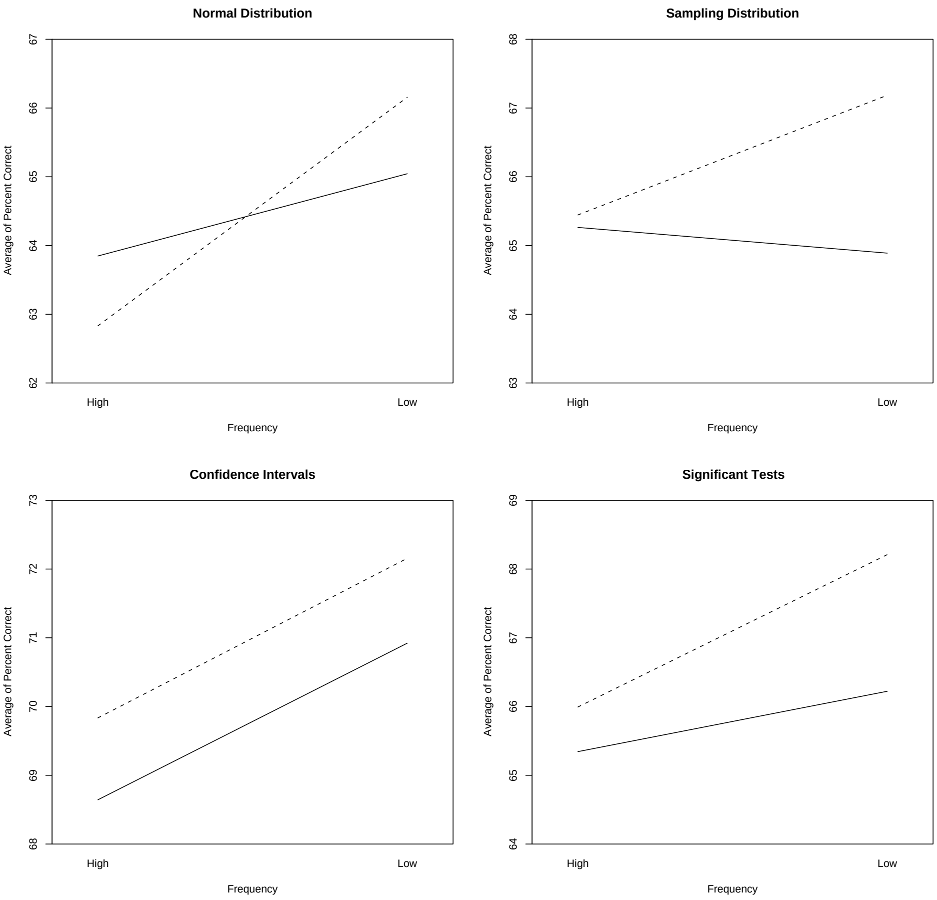

Figure 3.9: Average Percent Correct for the ARTIST Topic Scales. The solid line corresponds to Clustering *On*, the dashed line to Clustering *Off*.

learning outcomes. As can be seen from the first column of Table 3.6.1, the main effect of Frequency on engagement was estimated to be positive—indicating that asking more than 6 clicker questions is better than asking 3–4 questions—for each attitudinal outcome but was never significant at the 5% level. For each of the five subscales of the attitude survey, the estimated magnitudes of this effect were less than one-tenth of a percent (on a five point scale). The main effect of Frequency on learning, however, was negative. The estimated magnitude of this effect 1.4 percent (0.56 points on the 40-point scale for CAOS) and was non-significant at the 5% level.

There are several possible explanations for these results, the simplest being that

Table 3.26: Summary of Effects of Design Factors on Learning and Engagement

	Frequency		Clustering		Interaction	
	Estimate	Std.Error	Estimate	Std.Error	Estimate	Std.Error
Emotional Engagement						
Affect Subscale	0.04	0.02	-0.07	0.03	0.00	0.03
Value Subscale	0.02	0.02	0.00	0.02	0.02	0.02
Cognitive Engagement						
Cognitive Competence Subscale	0.01	0.02	-0.07	0.03	-0.03	0.02
Effort Subscale	0.00	0.03	-0.05	0.03	0.05	0.03
Attitude Toward Clickers						
Clickers Subscale	0.05	0.03	0.00	0.03	0.07	0.03
Learning						
CAOS	-1.37	0.39	1.60	0.45	-1.49	0.41

there is no effect of the number of clicker questions asked on engagement or learning. However, it is also possible that these results reflect limits in the design of this treatment factor. Recall from the description of the experimental design that the clicker questions were based on existing questions in the lab workbook, which contains activities that are more procedural in nature. Additionally, all lab sections were asked the same number of questions, with the same possible answer choices; the treatment groups differed with respect to the number of questions asked *with clickers*. Therefore, it is possible that:

1. There may have been a misalignment between the focus of the clicker questions and that of the CAOS and topic scale questions. The CAOS and topic scale questions were specifically written to capture students' conceptual understanding of Statistics, but many of the clicker questions were more factual in nature. This was due in part to the very purpose of the lab sections (and the questions in the lab workbook)—to reinforce and check understanding of concepts presented during lecture.

2. The differences between the treatment groups may have been too subtle to measure, since all sections were asked the same overall number of questions and differed only with respect to the physical clicking of the remote and display of the students' responses in bar-graph form.

3. Alternatively, there may have been too many questions at the *High* level, resulting in a general decrease in question quality. In particular, when more clicker questions were asked, there tended to be a higher proportion of quick check/recall questions (i.e. Do you remember that definition/rule?). Students may not have perceived much value in these questions and correspondingly provided lower ratings for those questions on the attitude survey pertaining to clickers. Indeed, the ratings for the emotional engagement sub-scale of the attitude survey, which contained questions specific to clickers, were lower than for the cognitive engagement sub-scale, which did not have any questions pertaining to clickers.

3.6.2 Discussion of RQ2

From the second column of Table 3.6.1, the main effect of Clustering on engagement was estimated to be negative for both subscales measuring cognitive engagement, as well as the Affect subscale measuring emotional engagement. However these effects were small and non-significant, each less than one-tenth of a percent (on a five point scale). For the Value and Clickers subscales, the estimated effect was nearly zero. The main effect of Clustering on learning was estimated to be positive—indicating that incorporating clicker questions throughout a class is better than asking them consecutively. The effects of Clustering on learning were larger than the effects of Clustering on engagement—1.6 percent (0.64 points on the 40-point scale for CAOS). While this effect was not significant at the 5% level, it was marginally significant at the 10% level. Additionally, the plots of performance on individual CAOS questions showed that the Blue Team, and to a lesser extent the Yellow Team (both with Clustering = *Off*), tended to outperform the teams where Clustering =

On. The Blue Team also outperformed the other teams for several CAOS topics. This provides some evidence that incorporating clicker questions throughout a class led to an increase in learning.

Logistically, it can be simpler to ask all clicker questions in a row, but the results of this experiment seem to imply that this may not benefit the students' understanding. This could be due in part to the position of the clicker questions within the material. Specifically, when clicker questions were clustered together during a lab session, they tended to come at the end of the lesson as a wrap-up, to review the concepts covered. Pedagogically, this could be useful to both student and instructor to see if the day's important points had been understood; there were several reports of this type of clicker use in the literature. However, this could change the cognitive level of a question and, correspondingly, the students' perceived value of the question. For example, a question asked before a topic is introduced could require students to apply existing knowledge to a new situation—extending their understanding—while the same questions asked after discussion of the topic could require students simply to remember what they had been told [12].

3.6.3 Discussion of H1

Looking at the final column of Table 3.6.1, the effect of the interaction between Frequency and Clustering was estimated to be positive—in contrast to the hypothesis—for four of the five attitudinal subscales, but the magnitudes of these effects were extremely small and non-significant at the 10% level. The effect of the interaction on learning was estimated to be negative. The magnitude of this effect was 1.49 percent (0.60 points on the 40-point CAOS scale) and was significant at the 10% level. In addition to this, several plots of the mean response, for both engagement and learning, by treatment factor did show descriptive evidence of interaction. All of this provides some evidence (albeit more qualitative than quantitative) for the hypothesis that asking too many clicker questions consecutively is not conducive to engagement nor to learning. Again, it is possible that limitations of the design factors affected the ability to measure this interaction. Refining and re-implementing this experiment may help shed light on the true effect of the interaction between Frequency and Clustering.

3.6.4 Discussion of RQ3.

Table 3.6.4 shows the number of additional students estimated to have used clickers under the *Moderate* and *High* levels of External Incentive as compared to the *Low* level, both when clicker use was defined as answering at least one clicker question and when it was defined as answering at least 50% of the clicker questions. Figure 3.10 shows the proportion of students using clickers for each level of External Incentive, collapsing over sequence and week. Based on these, it can be seen that clicker use significantly increases as the level of External Incentive increases. While this result is not necessarily surprising, it is somewhat disappointing. Previous studies have consistently indicated (based on student self-report) that students enjoyed using clickers and perceived some benefit, in terms of engagement and even learning, to their use. For the current experiment, it was hoped that this perceived value would affect students behavior, motivating them to use clickers even when there was no (or little) external influence to do so. However, this data does not support the idea that students perceived some inherent value to the clickers, at least not enough to affect their use of clickers. Even for those students who were required to use them early in the semester, and thus would have experienced their benefits, there was a decline in clicker use once it was no longer required (see Figures 3.4 and 3.5).

Table 3.27: Summary of Effects of External Incentive on Behavioral Engagement

Clicker Use	External Incentive			
	Moderate		High	
	Estimate	Std.Error	Estimate	Std.Error
At least one clicker question	0.751	0.299	1.792	0.332
At least 50% of the clicker questions	1.275	0.351	2.347	0.390

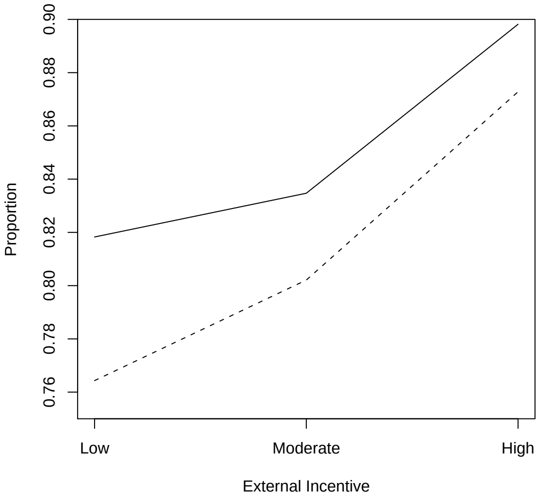

Figure 3.10: Proportion of Students Using Clickers by Level of External Incentive. The solid represents the proportion of students who answered at least one clicker question; the dashed line represents the proportion of students who answered at least 50% of the clicker questions.

3.7 Discussion: Experimental Design and Procedures

3.7.1 What went well in the experiment

A primary concern in designing this experiment was to ensure that experimental procedures were not too obtrusive or disruptive of normal class procedures. This is important for several reasons. As educators, our first responsibility is to our students, and we would not want an experiment that was detrimental to their learning experience or made them feel like experimental "guinea pigs." As researchers, we are under the governance of institutional review boards (IRB), which make sure that students' rights are protected. To achieve this end, several non-obtrusive elements were included in the design and implementation of this experiment.

First, the clicker questions were taken directly from existing questions in the students' lab workbook, so that no extra material was added into all ready full lab periods. Using questions that would have been asked anyway ensures that clicker use was seamlessly integrated into labs, increasing the intrinsic value of the questions and the clickers (i.e. clicker use is a part of the course, not something additional that students do not have to take seriously). Using existing questions also allowed for the same questions to be asked in every lab section, so that all students had access to the same material—a requirement of the University's IRB. Labs differed only with respect to the conditions under which clickers were used (e.g. the number of questions asked *with clickers*), correcting a design flaw in several previous clicker studies where clicker use was confounded with more general incorporation of active learning strategies (see Section 3.2.2 for further discussion on this).

Additionally, the instruments used to assess student learning were selected to provide formative feedback to students as well as summative assessment for the purposes of the experiment. Great care was taken to select instruments that could that could not only provide early feedback to students on their level of understanding (before losing points on homework assignments or exams), but also help *increase* their broad conceptual understanding. Use of nationally available, validated instruments corrected a common limitation of research in education: These instruments can be fully known by other researchers, helping frame the context of the results of the present research (e.g. Do higher scores indicate better conceptual understanding or better procedural ability?), as well as allowing for easier comparison of results across studies using the same outcome and easier reproduction of experimental conditions in future studies (see Section 2.3.2 for further discussion).

Finally, class time was provided for most assessments to be completed, placing minimal burden on the students' out-of-class time and ensuring higher completion rates for experimental activities. Often assessments were completed at the beginning of lab, hopefully decreasing the urge to rush through just to get it over with and get out the door. Extra credit—which is not typically available in the course—was offered as an incentive to complete two assessments that were not offered during class time.

Taking these steps lead to success on several fronts. First, it increased the lead course instructor's and several GSIs' ease with the experiment in general. They were all concerned about the well-being of the students, particularly that an equitable learning experience was maintained. GSIs were also concerned about their own workload, as they were taking several courses of their own. GSIs have specific workload restrictions negotiated and enforced by the University of Michigan's Graduate Employees Organization (GEO); care had to be taken to not exceed these restrictions. It also made seeking IRB approval easier, since the same workload and handling was guaranteed for every student in the course—those who provided consent to participate and those who did not. By making the experiment an integral part of the course, we did not have to seek student consent to participate in experimental procedures. Instead we sought student consent simply to use the data we collected for analysis and publication (students were ensured that their data would be used anonymously). Indeed, having well-planned and considerate experimental procedures lead to an extremely high consent rate—nearly 94% of the students enrolled in the course after the drop/add deadline agreed to participate.

Another successful, truly pragmatic, aspect of this experiment was the process for data collection and management. Having an easy-to-use, accurate system would be important in any experiment, but it becomes especially critical in such a large, complex experiment as this. Most data was collected using the online University of Michigan software called UM.Lessons. This software allowed data for every student in the class to be collected (and scored, when applicable) in one central database—without any data entry on the part of the researcher. This database was password protected, so only students enrolled in the course had access. Additionally, access could be set for certain days and times for students to complete the assessments, as well as view the questions or correct answers after submission. Data was securely backed-up on University servers and could be outputted in several formats for exploration and analysis. While the UM.Lessons service is specific to the University of Michigan, similar services may be available at other Universities. If not, there are several online data collection services available commercially. Certainly the use of online data collection is not new, but it is worth noting that implementation of this experiment really would not have been feasible without it.

3.7.2 What did not go well in the experiment

Despite best-efforts in planning, things are bound to go wrong (or at least not according to plan) in any experiment. For example, while great care was taken to select assessment instruments that would be beneficial to students, these instruments may not have been used to their fullest potential. Students completed COAS four times throughout the semester—prior to the start of treatment, the weeks near each of the two midterm exams, and after the completion of treatment. While the 40 questions and their answer choices were randomized each time, and students were at no point provided with solutions, students still became too familiar with the questions. It is possible that students did not take the final administration of CAOS seriously—answering based on familiarity rather than knowledge. Indeed, during mid-semester conversations, some students indicated that the questions were repetitive. So, as an incentive to increase effort in completing the final CAOS, students were awarded extra credit on their final exam for answering every question—and at least 50% of questions correctly. Additionally, feedback from students revealed that they did not consider the questions on CAOS or the in-lab reviews to be in line with questions on the course homework or exams. While the instruments were chosen specifically for their focus on conceptual issues—something that students often struggle with—many homework and several exam questions were problem-solving or procedurally based. A few of the CAOS questions were included on exams, but not enough to make students value use of this instrument in class. If this experiment were repeated, it would be better to have students complete CAOS only twice (pre- and post treatment) and would incorporate more of the CAOS and in-lab review questions directly on homework and exams. It could even be possible to have entire assessments take the place of a few standard homework assignments or a portion of the final exam, to increase their impact on course grades and thus naturally increase students' incentive to take them seriously.

As discussed in Section 3.3.1, there were inconsistencies in the implementation of each treatment factor. Variations in the number of clicker questions asked was often due to technical difficulties, which often cannot be controlled. Of greater concern was variation in the specific placement of individual clicker questions within a class period, since these variations could affect the cognitive level of the question. It would have been better for the integrity of this experiment to provide plans for each treatment group detailing exactly which questions were to be asked when and offering some scripted material for setting-up and debriefing questions. However, this would have been procedurally prohibitive, both in terms of time to develop such plans for four treatment groups over nine weeks, and in terms of excessive reduction of GSI freedom in teaching. In conversations with GSIs after the conclusion of the experiment, it was suggested that an alternative experimental procedure would be to manipulate clicker use during only a few weeks during the term, which might then make more extensive scripting and GSI training feasible. Also of concern were mistakes in the implementation of the crossover sequences. Several weeks were run at the wrong level of External Incentive. Additionally, on occasion, GSIs forgot to announce the crossover condition to students. If this experiment were repeated, more emphasize would need to be placed the pivotal switch points, as well as the weekly announcements of crossover condition to students.

Finally, it is possible that the very choice to implement the treatment in labs rather than lectures had serious, unintended consequences on the outcomes of the experiment. As has been mentioned before, lab sections were more plentiful in number and more uniform in terms of size and material taught. (While all lectures covered the same material using the same lecture notes, differences in their weekly scheduled meeting times resulted in differences in the timetable for covering the material.) The consistent schedule of lab once a week for 1.5 hours, with the exact same activities covered in each section, was much more conducive to the implementation of a factorial design. However, the very purpose of lab is to reinforce concepts presented during lecture. As a result, the clicker questions tended to be of lower cognitive value—focusing on recall or basic application, for example—thus reducing the need for deep thought on the part of the student to answer the question. Ultimately, this likely reduced the engagement and learning benefits of the clicker questions. This, in fact, could explain why there were so few significant results in the analyses.

3.8 Brief Summary and Overall Conclusions

This chapter presented the design and analysis of a experiment on the use of clickers in an introductory statistics course. The experiment had two main designs, run concurrently:

1. A two-factor design was used to explore the effects that the number of questions asked during a class period (Frequency) and the way those questions were incorporated into the material (Clustering) had on emotional and cognitive engagement as well as on learning.
2. A crossover design was used to explore the effect that grading or monitoring clicker use (External Incentive) had on behavioral engagement, as measured by the number of students who chose to use clickers.

Several hierarchical linear models of both engagement and learning outcomes were fit. Based on these analyses, there was little evidence that clicker use increased students' engagement, either emotionally, cognitively, or behaviorally. There was some evidence, however, that clicker use improved students learning. Increases in learning seemed to take place when the clicker questions were well incorporated into the material, particularly if the number of questions asked was low.

Taken together, the findings of this experiment provide a cautionary note for the educator interested in using clickers: As with any new technology or pedagogical technique, clickers may not be successful if they are not used in a well-planned, purposeful manner. The mere presence of clickers does not seem to be enough to engage students and thus improve learning. While the instant visual display of feedback from these devices is unique, it may not be valuable to students if the questions are poorly constructed.

CHAPTER IV

Exploiting interactions between educational interventions and ‘noise’ variables

4.1 Introduction

In this chapter, we discuss how to exploit interactions between *design factors* (educational interventions) and uncontrollable *noise variables* (e.g. student and instructor characteristics, classroom environments) to achieve two objectives: 1) Choose the settings of the educational interventions to reduce the sensitivity of the intervention to the noise variables, and 2) Choose the settings of the educational interventions to maximize their effects for subgroups of students and/or instructors.

The first approach has been used extensively in industries, especially manufacturing industries. It is referred to as robust design and was popularized by the Japanese quality consultant G. Taguchi [see 82, 89, 106, 116]. In the manufacturing context, robust design uses planned experiments to improve the design of products or manufacturing processes. In traditional approach to design of experiments, it is commonly assumed that the variance of the response is constant (or at most varies with the mean in a known way). In practice, however, both the means and variances depend on the input parameters (i.e. the design factors). The variance can be attributed to variations in the manufacturing, customer use or environmental conditions. The idea in robust design is to identify the important noise variables explicitly up front (in off-line experiments), vary them systematically, and find the influence of the design factors on both the mean response as well as the variance over the settings of the noise factors. Then, we choose the settings of the design factors so that we have as small variance as possible while also getting as close as possible to the desired response level on the average. There is an extensive literature on this topic, and a modern treatment can be found in Nair [82] and Wu and Hamada [116].

In many situations, however, the noise variables cannot be controlled even in off-line experiments. In such case, Freeny and Nair [41] proposed an approach where the values of the noise variables can be measured, and their interactions between the design factors and noise variables can be exploited to achieve robustness. This is the approach we will take in this chapter. In the educational context, this would be especially important for noise variables that are heterogeneous within a classroom—those variables that represent characteristics of the students themselves.

In the second objective, we do not want to select the design factor settings to mitigate the effect of noise factors. Rather, we want to use the information in the interaction to proactively customize the intervention to particular groups. In this case, we are tailoring the treatment to groups. In the educational context, this would be best for noise variables that are homogeneous within a classroom—those variables that represent class or instructor characteristics.

4.2 Simulated Example and Data

The response Y for this example is the level of statistical knowledge obtained by a student at the end of an introductory statistics course. Y is measured by score on the post treatment Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) instrument [31]. The treatments of interest represent combinations of three design factors, each with two levels:

A. Use of computer applets to demonstrate concepts

Low: Not used

High: Used

B. Use of clickers

Low: Not used

High: Used

C. Type of homework questions

Low: Basic applied repetition

High: Open-ended problem solving

Table 4.1: The design matrix

	Factor		
Run (i)	A	B	C
1	-1	-1	-1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	-1
5	-1	-1	+1
6	+1	-1	+1
7	-1	+1	+1
8	+1	+1	+1

We use the 2^3 full factorial design in Table 4.2 to study the resulting eight treatment combinations (also called *runs*). Each row represents a possible treatment combination. For each factor, -1 indicates the low level of that factor, while +1 indicates the high level. Since this is a full factorial design, we can study the individual effects of each factor and also investigate all interactions. We use eight *replicates*—i.e. we implement each of the treatment combinations in eight different classrooms. With eight treatment combinations and eight replicates, we need 64 classrooms to fully implement this experiment.

There are many potential sources of noise that could affect a student's response to the treatments, or that could affect our ability to detect true treatment effects. In this example, we consider seven noise variables:

- **n1**: The size of the class,
- **n2**: The time of day at which the class starts,
- **n3**: The instructor's attitude toward reform-oriented teaching,
- **n4**: The instructor's teaching experience,
- **n5**: A student's baseline knowledge of statistics,
- **n6**: A student's attitude toward statistics, and
- **n7**: A student's general scholastic aptitude.

We use the notation n_{tijs} to represent the value of the noise variable for student s in a class assigned to treatment combination i and replicate j. (For noise variables measured at the classroom-level (e.g. n_1 and n_2) or instructor-level (e.g. n_3 and n_4), the subscript s is dropped.) Recall that there is only one classroom assigned to each combination of run and replicate, for a total of 64 classrooms participating in the experiment. The total number of students is 3,055.

n1 Class size was coded as 1 ("Large") if a section's enrollment is more than 50 students and coded as -1 ("Small") otherwise. Half of the participating classes were large, accounting for 75% of the students in the sample.

n2 Class start time was coded as 1 if a section starts during the "Prime" hours of the day—namely, 11am to 3pm—and coded as -1 ("Off-prime") otherwise. Half of the start times were prime, accounting for 53% of the students in the sample.

n3 Instructor's attitude toward reform-oriented teaching was measured by the Faculty Attitude Towards Statistics (FATS) [2]. This instrument has twenty-five statements that are rated on a 5-point scale. Instructor ratings greater than 85 points on the FATS were coded as 1, indicating a “Good” attitude toward reform-oriented teaching; rating of 85 points or less were coded as -1, indicating a “Poor” attitude. Thirty-eight of the 64 instructors had good attitudes, accounting for 62% of the students in the sample.

n4 Instructor experience was measured as years teaching any class at the university level. It was coded as 1 if the instructor had been teaching for at least three years, indicating “High” experience and coded as -1 otherwise (“Low” experience). Thirty-five of the instructors had high teaching experience, accounting for 59% of the students in the sample.

n5 A student’s baseline knowledge of statistics was measured (on a continuous scale) by their pretreatment score on CAOS. Individual pretreatment CAOS scores were centered at the overall mean of 21 points.

n6 A student’s attitude toward statistics was measured by the Survey of Attitudes Toward Statistics (SATS) [97]. This instrument has thirty-six statements rated on a 7-point scale. Student ratings greater than 144 points on the SATS were coded as 1, indicating a “Good” attitude toward statistics; ratings of 144 points or less were coded as -1, indicating a “Poor” attitude. 70% of the students in the sample had good attitudes.

n7 A student’s general scholastic aptitude was measured by their grade point average (GPA) prior to the start of the experiment. GPA was coded as 1 is the student’s GPA was average or above (at least 3.00 points; “High”) and coded as -1 (“Low”) otherwise. 46% of the students in the sample had a high GPA.

The response of interest is Y_{ijs}, the post treatment CAOS score for student s in a class assigned to treatment combination i and replicate j. Table 4.2 shows the average of the centered post treatment scores for all students in a class. This table also shows the mean and standard deviation for each run of the experiment. For example, the average post treatment CAOS score for students in the second class ($i = 1, j = 2$) is 2.65 points above the average post treatment score.

Table 4.2: Average centered post treatment CAOS score

	Replicate (j)									
Run (i)	1	2	3	4	5	6	7	8	$\bar{Y}_i$	s_i
1	0.05	2.65	-0.35	-0.11	-2.39	0.54	0.54	0.25	0.00	3.86
2	-1.57	1.94	-2.97	4.90	-0.15	1.05	1.31	3.95	0.39	4.16
3	-1.35	4.19	0.15	4.72	0.42	0.65	3.01	-0.95	1.18	3.96
4	-2.01	1.80	0.97	-0.75	-1.01	-0.38	4.47	-0.11	0.17	3.60
5	2.42	2.29	0.49	1.55	1.44	-3.25	1.80	-1.58	0.00	3.91
6	0.57	-2.69	-3.83	-2.96	3.70	1.30	-1.42	4.90	-1.00	4.32
7	-0.80	-0.42	4.03	1.58	-1.14	1.85	-0.23	4.94	1.52	4.34
8	-1.04	-2.23	-3.88	-2.61	-0.27	-0.61	1.24	-1.05	-1.60	3.53

4.3 A General Model to Account for Multiple Noise Variables

In general, suppose that we model the response for student s in a class assigned to treatment combination i and replicate j as:

$$Y_{ijs} = \boldsymbol{x_i'\alpha} + \sum_{t=1}^{r} (\boldsymbol{x_i'\phi_t}) n_{tijs} + \epsilon_{ij} + \delta_{ijs}, \tag{4.1}$$

where x_i represents the i^{th} row of the design matrix (the i^{th} treatment combination) for $i = 1, \ldots, m$, $j = 1, \ldots, k$ corresponds to a repetition of that treatment combination, and $t = 1, \ldots, r$ corresponds to the number of noise variables. Here $\boldsymbol{\alpha}$ represents a vector of *location effects*, or the effects of the design factors on the average value of the response. There are r vectors $\boldsymbol{\phi_t}$—one for each noise variable—that represent the effect of the interaction between noise variable $\boldsymbol{n_t}$ and the design factors on the response. The $\boldsymbol{\phi_t}$ are referred to as *dispersion effects*.

We seek to exploit this functional relationship to 1) identify particular settings of the design factors for which the variation in the response due to each noise variable is minimal, or 2) identify particular groups to which we should tailor treatment. The steps of an analytic strategy to accomplish these goals are given within the context of the statistics education example presented in the previous section.

4.4 Illustrating the Analysis Strategy to Exploit Interactions

Using the data described in Section 4.5, the analytic strategy will be implemented in five steps.

Step 1: Determine the appropriate functional relationship between the response and each continuous noise variable.

The functional functional relationship between the response and each continuous noise variable could be determined based on prior knowledge, or it could be determined graphically based on data from the experiment. It this example, the only continuous noise variable is the student's baseline knowledge of statistics, as measured by their pretreatment CAOS score. Figure 4.1 plots the relationship between the pre and post treatment scores for each treatment combination. Least-square linear regression lines are superimposed and appear to be a good fit for this data, indicating that a linear relationship between pre and post treatment CAOS scores is reasonable.

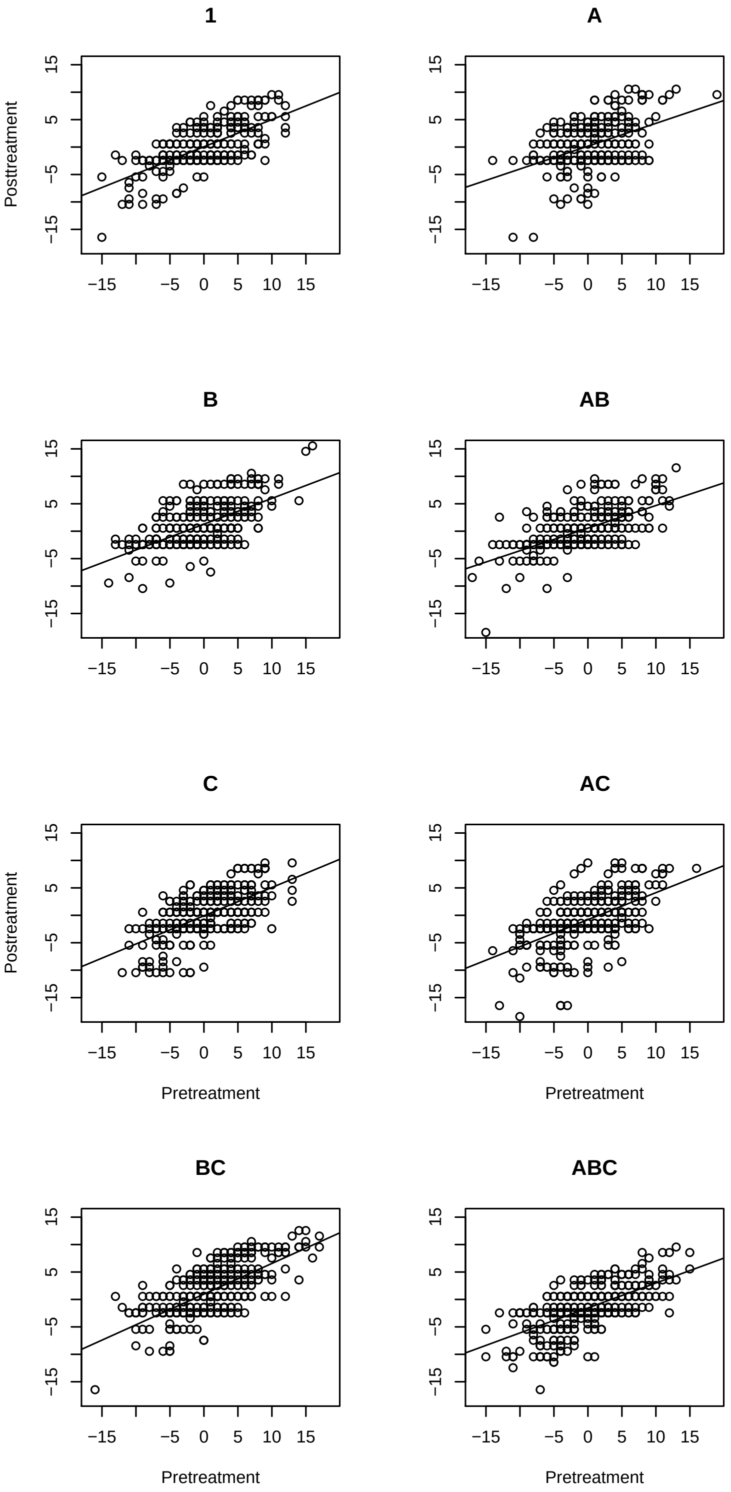

Figure 4.1: Post Treatment vs. Pretreatment CAOS Scores for the Eight Treatment Combinations. These plots are used to suggest the functional relationship between the response (post treatment scores) and the continuous noise variable (pretreatment scores). Least-square linear regression lines are superimposed, and appear to be a good fit for each plot. This indicates that a liner relationship between pre and post treatment CAOS scores is reasonable.

Step 2A: Obtain initial estimates of the location effects $\boldsymbol{\alpha}$ and dispersion effects $\boldsymbol{\phi_t}$.

Initial estimates of all location and dispersion effects are obtained by fitting the full model (4.1) for the response. Since the data in example represents students who are nested within classrooms, a hierarchical linear model is fit with random effects for each of the 64 classes. Table 4.4 presents partial results from this model; due to the length of the output, only those effects with p-values less than 0.200 are shown. Estimated effects that are significant at the 5% level are shown in bold. These significant effects will be used in Step 2B to refine model for the response.

Table 4.3: Initial estimates of location and dispersion effects

	Estimate	Std.Error	DF	p-value
Intercept	0.025	0.223	2932	0.912
$\boldsymbol{A}$	1.056	0.125	2932	**0.000**
$\boldsymbol{B}$	0.421	0.123	2932	**0.001**
$\boldsymbol{n_1}$	-0.944	0.212	59	**0.000**
$\boldsymbol{n_3}$	0.890	0.218	59	**0.000**
n_4	0.359	0.212	59	0.095
$\boldsymbol{n_5}$	0.490	0.007	2932	**0.000**
$\boldsymbol{n_6}$	0.985	0.035	2932	**0.000**
$\boldsymbol{n_7}$	0.514	0.032	2932	**0.000**
A:B	0.206	0.116	2932	0.076
$\boldsymbol{B}$:$\boldsymbol{n_1}$	0.436	0.100	2932	**0.000**
C:n_1	0.192	0.112	2932	0.087
$\boldsymbol{A}$:$\boldsymbol{n_3}$	0.185	0.092	2932	**0.044**
$\boldsymbol{B}$:$\boldsymbol{n_3}$	0.542	0.094	2932	**0.000**
$\boldsymbol{A}$:$\boldsymbol{n_4}$	-0.385	0.089	2932	**0.000**
$\boldsymbol{C}$:$\boldsymbol{n_4}$	-0.174	0.077	2932	**0.024**
B:n_5	0.009	0.007	2932	0.160
$\boldsymbol{A}$:$\boldsymbol{n_6}$	-0.506	0.035	2932	**0.000**
$\boldsymbol{B}$:$\boldsymbol{n_6}$	0.488	0.035	2932	**0.000**
A:n_7	-0.061	0.032	2932	0.062
A:B:n_1	-0.152	0.106	2932	0.151
$\boldsymbol{A}$:$\boldsymbol{C}$:$\boldsymbol{n_1}$	0.206	0.099	2932	**0.038**
A:B:n_2	0.136	0.090	2932	0.131
A:C:n_6	-0.048	0.035	2932	0.171
$\boldsymbol{A}$:$\boldsymbol{B}$:$\boldsymbol{n_7}$	0.070	0.033	2932	**0.032**
$\boldsymbol{B}$:$\boldsymbol{C}$:$\boldsymbol{n_7}$	-0.077	0.032	2932	**0.018**

Step 2B: Refine the model for the response.

Those effects estimated to be significant at the 5% level in Step 2A are fit in a reduced model for the response. In the cases where an interaction was significant, the corresponding main effects were also included even if they were not individually significant (this is referred to as the "hierarchy principle" in experimental design).

For example, the interaction B:C:n_7 was significant at the 5% level, so each of the main effects for B, C, and n_7 were also included in the reduced model. Results for the reduced model are shown in Table 4.4. Again, effects which are estimated to be significant at the 5% level are in bold. This model can be refined further—the effects A:n_4, C:n_4, and A:C:n_1 can be removed without violating the hierarchy principle. After removing these terms, no further reductions can be made. The final model for the response is presented in Step 3.

Table 4.4: Refined estimates of location and dispersion effects

	Estimate	Std.Error	DF	p-value
Intercept	-0.116	0.204	2972	0.571
$\boldsymbol{A}$	1.107	0.069	2972	**0.000**
$\boldsymbol{B}$	0.373	0.077	2972	**0.000**
C	0.044	0.053	2972	0.406
$\boldsymbol{n_1}$	-0.922	0.199	60	**0.000**
$\boldsymbol{n_3}$	0.965	0.203	60	**0.000**
n_4	0.352	0.200	60	0.085
$\boldsymbol{n_5}$	0.489	0.006	2972	**0.000**
$\boldsymbol{n_6}$	0.988	0.035	2972	**0.000**
$\boldsymbol{n_7}$	0.516	0.032	2972	**0.000**
A:n_1	-0.051	0.068	2972	0.449
$\boldsymbol{B}$:$\boldsymbol{n_1}$	0.471	0.071	2972	**0.000**
A:n_3	0.075	0.052	2972	0.147
$\boldsymbol{B}$:$\boldsymbol{n_3}$	0.566	0.057	2972	**0.000**
$\boldsymbol{A}$:$\boldsymbol{n_4}$	-0.388	0.052	2972	0.000
C:n_4	-0.057	0.052	2972	0.272
$\boldsymbol{A}$:$\boldsymbol{n_6}$	-0.512	0.035	2972	**0.000**
$\boldsymbol{B}$:$\boldsymbol{n_6}$	0.485	0.035	2972	**0.000**
A:n_7	-0.057	0.032	2972	0.075
B:n_7	0.007	0.032	2972	0.839
C:n_7	-0.003	0.032	2972	0.921
$\boldsymbol{A}$:$\boldsymbol{B}$:$\boldsymbol{n_7}$	0.069	0.032	2972	**0.033**
$\boldsymbol{B}$:$\boldsymbol{C}$:$\boldsymbol{n_7}$	-0.076	0.032	2972	**0.019**

Step 3: Estimate the model for the response based on the active location and dispersion effects identified in Step 2.

Now, using the final reduced model identified in Step 2B, we re-estimate the parameter values and compute fitted values for the response. The final model for the response is presented in Table 4.4. The parameter estimates from this model will be used in Step 4 to select settings of the design factors for which the response is maximized but the dispersion effects are minimized, or cases where settings of the design factors should be customized.

Table 4.5: Final estimates of location and dispersion effects

	Estimate	Std.Error	DF	p-value
Intercept	-0.142	0.204	2975	0.488
$\boldsymbol{A}$	1.078	0.051	2975	**0.000**
$\boldsymbol{B}$	0.381	0.077	2975	**0.000**
C	0.031	0.050	2975	0.530
$\boldsymbol{n_1}$	-0.923	0.200	60	**0.000**
$\boldsymbol{n_3}$	0.967	0.204	60	**0.000**
n_4	0.348	0.201	60	0.089
$\boldsymbol{n_5}$	0.489	0.006	2975	**0.000**
$\boldsymbol{n_6}$	0.990	0.035	2975	**0.000**
$\boldsymbol{n_7}$	0.516	0.032	2975	**0.000**
$\boldsymbol{B}$:$\boldsymbol{n_1}$	0.465	0.071	2975	**0.000**
$\boldsymbol{B}$:$\boldsymbol{n_3}$	0.568	0.055	2975	**0.000**
$\boldsymbol{A}$:$\boldsymbol{n_4}$	-0.359	0.049	2975	**0.000**
$\boldsymbol{A}$:$\boldsymbol{n_6}$	-0.513	0.035	2975	**0.000**
$\boldsymbol{B}$:$\boldsymbol{n_6}$	0.485	0.035	2975	**0.000**
A:n_7	-0.056	0.032	2975	0.078
B:n_7	0.007	0.032	2975	0.819
C:n_7	-0.004	0.032	2975	0.903
$\boldsymbol{A}$:$\boldsymbol{B}$:$\boldsymbol{n_7}$	0.068	0.032	2975	**0.034**
$\boldsymbol{B}$:$\boldsymbol{C}$:$\boldsymbol{n_7}$	-0.078	0.032	2975	**0.016**

Step 4: Determine improved settings of the design factors.

Table 4.4 shows the effects of the design factors on the average post treatment CAOS score. Figure 4.2 also shows these effects. Two things should be mentioned here before interpreting these specific effects. First, recall that this data was simulated; this example is meant to illustrate how this analytic strategy could be implemented in an educational setting, and the substantive conclusions are meant to be representative of the types of conclusions that could made when using this strategy. Second, the response was centered before fitting any regression models, so that a Y value of zero represents the average post treatment CAOS score.

Table 4.6: Estimated effects of design factors

Location Effects:					
Complete notes provided			Partial notes provided		
	Clickers			Clickers	
Applets	Not used	Used	Applets	Not used	Used
Not used	-1.632	-0.870	Not used	-1.570	-0.808
Used	0.524	1.286	Used	0.586	1.348

Dispersion Effects Due To:					
Class Size		**Instructor Attitude**		**Instructor Experience**	
Clickers		Clickers		Applets	
Not used	-1.388	Not used	0.399	Not used	0.707
Used	-0.046	Used	1.535	Used	-0.011

Student Attitude		
	Clickers	
Applets	Not used	Used
Not used	1.018	1.988
Used	-0.008	0.962

Student Grade Point Average					
Complete notes provided			Partial notes provided		
	Clickers			Clickers	
Applets	Not used	Used	Applets	Not used	Used
Not used	0.599	0.593	Not used	0.707	0.429
Used	0.311	0.617	Used	0.459	0.453

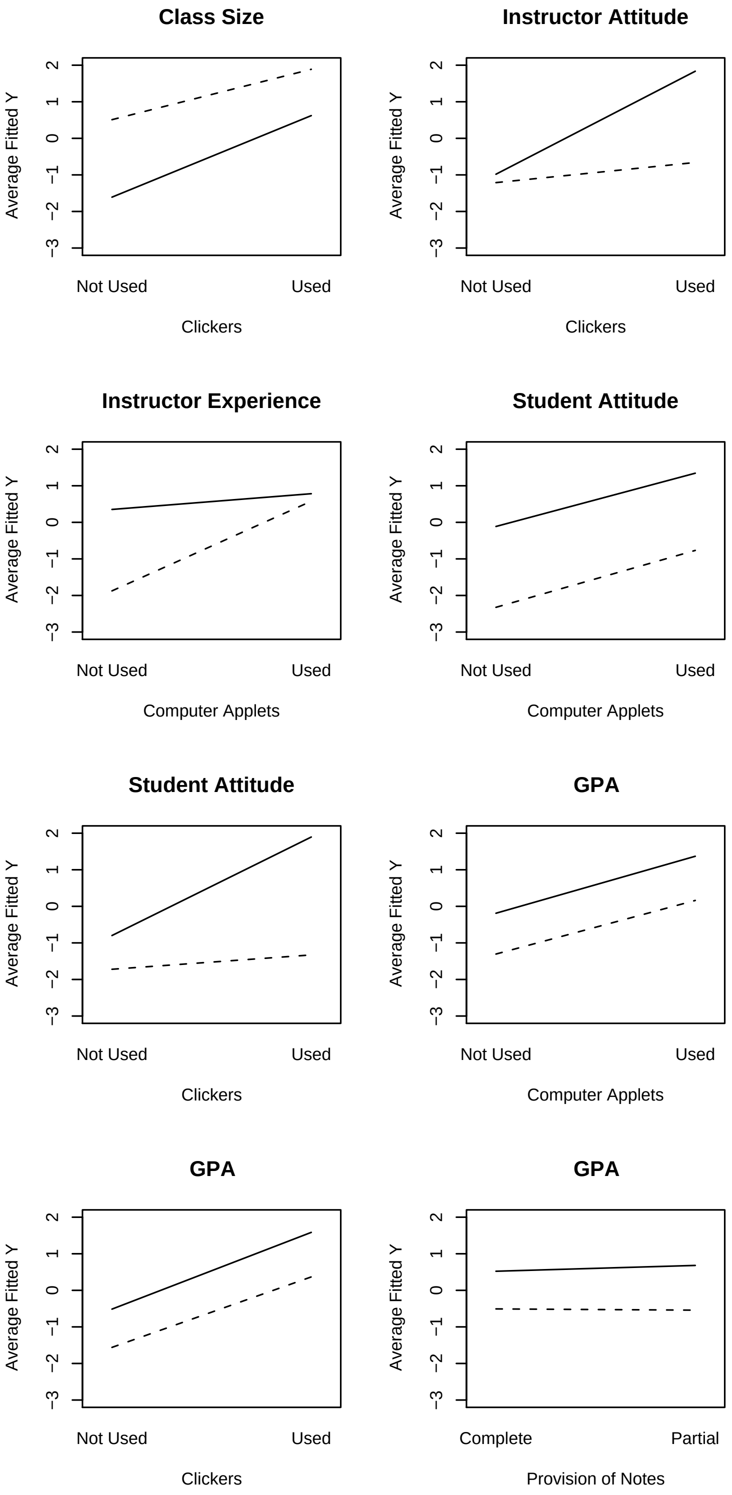

Figure 4.2: Effect of Interaction Between the Design Factors and Noise Variables on the Average Fitted Post Treatment CAOS Score. For each panel, the solid line corresponds to the high level of the noise variable given in the panel title; the dashed line corresponds to the low level.

Based on these effects, we could make the following conclusions:

- All of the design factors have an effect on post treatment CAOS scores, though the effect of provision of lecture notes was not significant at a 5% level. From Table 4.4, using applets results in a 1.078 point increase above the average posttreatment CAOS score while using clickers results in a 0.381 point increase. When evaluating the location effects, the goal is to find the treatment combination that maximizes the response. The largest total effect occurs when partial notes are provided and both applets and clickers are used—this will lead to a predicted 1.348 point increase in the average post treatment CAOS score (see "Location Effects" in Table 4.4).
- There is an interaction between class size and use of clickers. Dispersion effects due to class size could be minimized by using clickers (see the section of Table 4.4 and Figure 4.2 entitled "Class Size"). However, from the figure, it can be seen that the estimated gains in the response from using clickers is greater in large classes. Here is an instance where treatment could be tailored—a large course clearly benefits from the use of clickers, while a small class performs similarly regardless of clicker use. Other considerations, such as the cost of incorporating this technology, could influence the decision to use (or not use) clickers in a small class.
- There appear to be no effects of class start time on the average post treatment CAOS score, since it was not included in the final model for the response.
- There is an interaction between instructor attitude and use of clickers. Dispersion effects due to instructor attitude could be minimized by not using clickers, however, instructors with good attitudes have much to gain from clicker use. Again, it may be reasonable to tailor treatment here. An instructor who has

a favorably disposed reform-oriented teaching should consider using clickers, whereas an instructor who is not so favorably disposed might as well not use clickers.

- There is an interaction between instructor experience and use of applets. Dispersion effects due to instructor experience could be minimized by using applets (see “Instructor Experience” in Table 4.4). Additionally, post treatment CAOS scores are higher on average when applets are used regardless of whether the instructor has a high level of teaching experience or not (see corresponding section of Figure 4.2). Together, these provide support for using computer applets.
- There is no interaction between a student’s baseline knowledge of statistics and any of the design factors, indicating that changing the settings of the design factors cannot mitigate the effect of baseline knowledge on the student’s knowledge at the end of the course. From Table 4.4, the coefficient for n_5 is 0.489 points. For each point increase above the average pretreatment CAOS score, post treatment CAOS score is expected to increase nearly half a point above the post treatment average.
- There is an interaction between student attitude and the use of applets and also between student attitude and the use of clickers. Since we can expect heterogeneous student attitudes within a class, it would be better to find settings of the design factors which are robust to this noise variable, rather than to tailor treatment. Dispersion effects due to student attitude are minimized when applets are used but clickers are not (see Table 4.4, as well as the two panels entitled “Student Attitude” in Figure 4.2). In fact, the estimated effect of student attitude at these settings of the design factors is nearly zero (-0.008), indicating that the response is robust to changes in student attitude under this

treatment combination.

- While estimates of an interaction between student grade point average and the design factors were statistically significant at the 5% level (see Table 4.4, there appears to be little practical effect of this interaction on post treatment CAOS scores. This can be seen by this similar magnitude of the effects presented in Table 4.4, as well as the parallel lines in the three panels entitled “GPA” in Figure 4.2. It would seem that changing the settings of the design factors does not really mitigate the dispersion effects due to grade point average. Also, there does not seem to be a subgroup of students for whom it would make sense to customize treatment for any of the design factors.

After assessing the effects of the individual noise variables, the conclusions should be evaluated in light of the current understanding of each factor, as well as current theories on teaching and learning, to design an effective treatment for all students, or to identify case when it might make sense to customize treatment for a subgroup of students/instructors.

4.5 Summary and Discussion

This chapter illustrated, in an educational context, the application of a data analytic strategy that exploits interactions to identify the best treatment, either overall or for a particular subgroup. This strategy can be implemented in four steps:

Step 1 Determine the appropriate functional relationship between the response and each continuous noise variable.

Step 2A Obtain initial estimates of the location effects $\boldsymbol{\alpha}$ and dispersion effects $\boldsymbol{\phi_t}$.

Step 2B Refine the model for the response.

Step 3 Estimate the model for the response based on the active location and dispersion effects identified in Step 2.

Step 4 Determine improved settings of the design factors.

In general, studies that employ this strategy require one more replication than the number of noise variables to be studied. Given the numerous noise variables that could be present in educational data, most of these studies in education will need to be large, involving many classrooms. Replications could be accumulated through coordination between universities, by including courses of different levels or from different disciplines, or by repeating the experiment over time. Due to their size, these studies are best suited as well-planned follow-up studies. It will be important to identify several design factors for which effectiveness has already been demonstrated. It will also be important to identify those noise variables that are most likely to interact with treatment, affecting the response to that treatment. Candidate noise variables can be identified using expert opinion, current theory on teaching and learning, or through previous research. To minimize the differences between the classrooms in which the study is implemented, common assessment instruments and implementation procedures will need to be used in all classrooms, to the greatest extent possible. Remaining differences between classrooms could be included as noise variables to be studied. This could include characteristics of the course itself (e.g. level, meeting time, length, size), the instructor (attitude, experience), the institution (liberal arts college vs. research institution, rural vs. urban elementary school, quality of facilities), or even the degree of implementation fidelity. In fact, details pertaining to implementation should likely be key noise variables in these studies. This is because it is so difficult in educational settings to separate treatment from the mechanism through which it is delivered [e.g. 15, 84]. One area of future research that could increase the potential of these studies to impact education will be the systematic identification and quantification of treatment implementation—which aspects of implementation are truly important to measure and how to measure them. Once identified, characteristics of implementation could be divided into two groups: those that we would like to learn about explicitly, and those that we would like the response to be robust to. Characteristics of the first kind could be studied as design factors (e.g. whether it is better to ask a large or small number of clicker questions); characteristics of the second kind could be studies as noise variables (e.g. instructor enthusiasm toward treatment).

The unavoidable intertwining of treatment and its implementation in educational research relates to a major concern in this research field: The ability to attribute an improvement to the treatment itself, rather than the natural growth of students or other confounding factors. It also has direct implications on the ability to generalize findings from one research study to another group of students. For example, it has been noted that some studies which seek to determine the effectiveness of clickers by comparing a section where clicker questions are asked to a traditional lecture section are in fact measuring the effect of active learning strategies in general—clicker use is incidental [23]. As another example, suppose two studies report on the effectiveness of using computer applets to demonstrate concepts. In one study, students work on the applet in groups but with little guidance from the instructor; in the other study, student groups are given clear objectives to work toward and a final "wrap-up" of the concepts demonstrated. The results of each study would then reflect their implementation differences as much as the actual effectiveness of the treatment (if any). The primary consequence of this intertwining is that we cannot be sure of our ability to replicate the findings from one educational research study in other classrooms. The "gold standard" in establishing causality is random assignment, however this is difficult to achieve in educational settings. It is rare to randomize individual students (as indicated by the review in Chapter II), and randomization of self-selected groups does not afford baseline equivalence the same way the individual randomization can. An alternative method for establishing causality is to repeat a study over time in diverse settings—if similar results can be obtained, this will build support that they are due to the treatment itself rather than the nuances of implementation. This, however, requires many small studies over a long period of time. Additionally, the ability for a study to be properly replicated can be limited by inconsistencies in the reporting of study conditions (see Section 2.3.6). Use of this strategy is an improvement over this process, since all replications take place at the same time and the treatment protocol is know explicitly by all sites (the classrooms). While these studies would not be trivial to implement, the end result could be more generalizable research—successful treatments that can be reproduced in a broad array of classrooms.

Appendix: Data Simulation

Values for the seven noise variables were selected to plausible within an educational context. For example, it would seem reasonable to obtain a number of classes that start at prime or off-prime hours, since classes could start at the top or bottom of each hour from 8am to 6pm. Each of the dichotomous noise variables, as well as the continuous, centered measure of baseline knowledge of statistics, were used to generate the post treatment CAOS scores. The entire process of generating the response was completed in steps. First, the ability for student s in a classroom receiving treatment combination i and replicate j to answer CAOS question q was generated according to the model

$$Z_{ijsq} = \alpha_{Ai} + \alpha_{Bi} + (\phi_{10} + \phi_{1Bi})n_{1ij} + (\phi_{30} + \phi_{3Bi})n_{3_ij} + (\phi_{40} + \phi_{4Ai})n_{4ij}$$

$$+\phi_{50}n_{5ijs} + (\phi_{60} + \phi 6Ai + \phi_{6Bi})n_{6ijs} + \phi_{70}n_{7ijs} + \epsilon_{ij} + \delta_{ijs} + \nu_{ijsq}, \tag{4.2}$$

where ϵ_{ij}, δ_{ijs} and ν_{ijsq} were each generated to have a normal distribution with a mean of 0 and a standard deviation of 3. Next, a difficulty score for each CAOS question q was generated as

$$D_q = \Phi^{-1}_{E(Z),var(Z)}(d_q), \tag{4.3}$$

where d_q represents the percent of incorrect responses to question q (forty values for d_q were selected between 30% and 70%). Finally, the total post treatment CAOS score Y_{ijs} was calculated as the sum of the forty indicators that $Z_{ijsq} - D_q \geq 0$.

BIBLIOGRAPHY

Bibliography

[1] Guidelines for assessment and instruction in statistics education (GAISE) college report, 2005. www.amstat.org/Education/gaise/GAISECollege.htm.

[2] Development and validation of a scale for measuring instructors attitudes toward concept-based or reform-oriented teaching of introductory statistics in the health and behavioral sciences, 2007. Doctoral Dissertation available at http://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Hassad.Dissertation.pdf.

[3] The introductory statistics course: A ptolemaic curriculum?, 2007. http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1.

[4] Using statistics effectively in mathematics education research (SMER): A report from a series of workshops organized by the american statistical association, 2007. http://www.amstat.org/research_grants/pdfs/SMERReport.pdf.

[5] C.L. Aberson, D.E. Berger, M.R. Healy, and V.L. Romero. An interactive tutorial for teaching statistical power. *Journal of Statistics Education*, 10, 2002. http://www.amstat.org/publications/jse/v10n3/aberson.html.

[6] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[7] J.R. Alldredge and G.R. Brown. Association of course performance with student beliefs: An analysis by gender and instructional software environment. *Statistics Education Research Journal*, 5:64–77, 2006.

[8] J.R. Alldredge, H.D. Johnson, and J.J. Sanchez. Does viewing video of statistics in action affect student attitudes? In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[9] J.R. Alldredge and N.A. Som. Comparison of multimedia educational materials used in an introductory statistical methods course. In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics*. International Statistical Institute, 2002.

[10] R. Auras and L. Bix. Wake up! The effectiveness of a student response system in a large packaging class. *Packaging Technology and Science*, 20:183–195, 2007.

[11] P. Ayres and J. Way. The effectiveness of using a video-recording to reproduce randomly generated sequences in probability research. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics.* International Statistical Institute, 1998.

[12] I.D. Beatty. Transforming student learning with classroom communication systems. *Educause Center for Applied Research, Research Bulletin*, 2004:2–13, 2004.

[13] I.D. Beatty, W.J. Gerace, W.J. Leonard, and R.J. Dufresne. Designing effective questions for classroom response system technology. *American Journal of Physics*, 74:31–39, 2006.

[14] W. Beekes. The 'millionaire' method for encouraging participation. *Active Learning in Higher Education*, 7:25–36, 2006.

[15] D.C. Berliner. Educational research: the hardest science of all. *Educational researcher*, 31(8):18–20, 2002.

[16] M. Bijker, G. Wynants, and H. van Buuren. A comparative study of the effects of motivational and attitudinal factors on studying statistics. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics.* International Statistical Institute, 2006.

[17] M. Bolzan. An experiment in teaching statistics at primary school in Italy. In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics.* International Statistical Institute, 2002.

[18] C.A. Brewer. Near real-time assessment of student learning and understanding in biology courses. *Bioscience*, 54:1034–1039, 2004.

[19] P. Brickman. The case of the Druid Dracula: A directed "clicker" case study on DNA fingerprinting. *Journal of College Science Teaching*, 36:48–53, 2006.

[20] D.M. Bunce, J.R. VandenPlas, and K.L. Havanki. Comparing the effectiveness on student achievement of a student response system versus online WebCT quizzes. *Journal of Chemical Education*, 83:488–493, 2006.

[21] U. Bunz. Using scantron versus an audience response system for survey research: Does methodology matter when measuring computer-mediated communication competence? *Computers in Human Behavior*, 21:343–359, 2005.

[22] J.E. Caldwell. Clickers in the large classroom: Current research and best–practice tips. *CBE Life Sciences Education*, 6:9–20, 2007.

[23] C. Carnaghan and A. Webb. Investigating the effects of group response systems on student satisfaction, learning and engagement in accounting education, Dec 2006. Available: `http://ssrn.com/abstract=959370`.

[24] Y.F. Chen, C.C. Liu, M.H. Yu, S.B. Chang, Y.C. Lu, and T.W. Chan. Elementary science classroom learning with wireless response devices implementing active and experimental learning. In *Proceedings of the Third IEEE International Workshop on Wireless and Mobile Technologies in Education*, pages 96–103, Tokushima, Japan, Nov 2005.

[25] G. Cicchitelli and G. Galmacci. The impact of computer programs on the learning of descriptive statistics: The case of DSTATS. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics*. International Statistical Institute, 1998.

[26] A.M. Cleary. Using wireless response systems to replicate behavioral research findings in the classroom. *Teaching of Psychology*, 35, 2008.

[27] L.B. Collins and K.C. Mittag. Effect of calculator technology on student achievement in an introductory statistics course. *Statistics Education Research Journal*, 4:7–14, 2005.

[28] J. Conoley, G. Moore, B. Croom, and J. Flowers. A toy or a teaching tool? The use of audience-response systems in the classroom. *Journal of the Association for Career and Technical Education*, pages 46–48, 2006.

[29] C.H. Crouch and E. Mazur. Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69:970–977, 2001.

[30] N. Davies, R. Lees, S. Smith, and R. O'Neill. Integrating Microsoft Windows applications with modules for learning statistics, mathematics and other core curriculum areas. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics*. International Statistical Institute, 1998.

[31] R. delMas, J. Garfield, B. Chance, and A. Ooms. Assessing students' conceptual understanding after a first course in statistics. 2006. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

[32] C. Demetry. Use of educational technology to transform the 50-minute lecture?: Is student response dependent on learning style. In *Proceedings of the 2005 American Society for Engineering Education Annual Conference and Exposition*. American Society for Engineering Education, 2005.

[33] I. Dinov and J. Sanchez. Assessment of the pedagogical utilization of the statistics online computational resource in introductory probability courses: A quasi-experiment. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[34] D. Duncan. *Clickers in the Classroom: How to enhance science teaching using classroom response systems*. Pearson, San Fransisco, CA, 2005.

[35] J. Dutton and M. Dutton. Characteristics and performance of students in an online section of business statistics. *Journal of Statistics Education*, 13, 2005. www.amstat.org/publications/jse/v13n3/dutton.html.

[36] P.B. Elmore and P.L. Woehlke. Twenty years of research methods employed in American Education Research Journal, Educational Researcher, and Review of Educational Research. San Diego, CA, 1998. Paper presented at the Annual Meeting of the American Educational Research Association.

[37] F. Enders and M. Diener-West. Methods of learning in statistical education: A randomized trial of public health graduate students. *Statistics Education Research Journal*, 5:5–19, 2006.

[38] A.P. Fagan, C.H. Crouch, and E. Mazur. Peer instruction: Results from a range of classrooms. *The Physics Teacher*, 40:206–209, 2002.

[39] J.A. Fredricks, P.C. Blumenfeld, and A.H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74:59–109, 2004.

[40] S. Freeman, E. O'Connor, J.W. Parks, M. Cunningham, D. Hurley, D. Haak, C. Dirks, and M.P. Wenderoth. Prescribed active learning increases performance in introductory biology. *CBE Life Sciences Education*, 6:132–139, 2007.

[41] A.E. Freeny and V.N. Nair. Robust parameter design with uncontrolled noise variables. *Statistica Sinica*, 2(2):313–334, 1992.

[42] J. Garfield. How students learn statistics. *International Statistical Review*, 63:25–34, 1995.

[43] J. Garfield and A. Ahlgren. Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 19:44–63, 1988.

[44] J. Garfield and D. Ben-Zvi. How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75:372–396, 2007.

[45] O. Giambalvo, A.M. Milito, and A.M. Oliveri. The results of a performance test: A multilevel analysis. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[46] J.A. Gonzalez, L. Jover, E. Cobo, and P. Munoz. Formal assessment of an innovative web-based tool designed to improve student performance in statistics. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[47] L.D. Goodwin and W.L. Goodwin. An analysis of statistical techniques used in the Journal of Educational Psychology, 1979-1983. *Educational Psychologist*, 20:13–21, 1985.

[48] L.D. Goodwin and W.L. Goodwin. Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educations research literature. *Educational Researcher*, 14:5–11, 1985.

[49] L. Greer and P.J. Heaney. Real-time analysis of student comprehension: an assessment of electronic student response technology in an introductory earth science course. *Journal of Geoscience Education*, 52:345–351, 2004.

[50] P. Haidet, D. Hunt, and J. Coverdale. Learning by doing: Teaching critical appraisal of randomized trials be performing an in-class randomized trial. *Academic Medicine*, 77:1161–1162, 2002.

[51] J.T. Hanley and P. Jackson. Making it click: A California high school test drives and evaluates six new personal response systems. *Technology and Learning*, 26:34–38, 2006.

[52] RK Henson. Expanding reliability generalization: Confidence intervals and Charter's combined reliability coefficient. *PERCEPTUAL AND MOTOR SKILLS*, 99(3, Part 1):818–820, 2004.

[53] C.F. Herried. "Clicker" cases: Introducing case study teaching into large classrooms. *Journal of College Science Teaching*, 36:43–47, 2006.

[54] S.C. Hilton and H.B. Christensen. Evaluating the impact of multimedia lectures on student learning and attitudes. In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics*. International Statistical Institute, 2002.

[55] Ryan T. Howell and Alan L. Shields. The file drawer problem in reliability generalization - A strategy to compute a fail-safe N with reliability coefficients. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 68(1):120–128, 2008.

[56] S.R. Hutchinson and C.D. Lovell. A review of methodological characteristics of research published in key journals in higher education: Implications for graduate research training. *Research in Higher Education*, 45:383–403, 2004.

[57] E.H. Ip. Visualizing multiple regression. *Journal of Statistics Education*, 9, 2001. `www.amstat.org/publications/jse/v9n1/ip.html`.

[58] M.H. Jackson and A.R. Trees. Clicker implementation and assessment. Available: `comm.colorado.edu/mjackson/clickerreport.htm`.

[59] M.C. James. The effect of grading incentive on student discourse in peer instruction. *American Journal of Physics*, 74:689–691, 2006.

[60] D. Kaplan. Computing and introductory statistics. *Technology Innovations in Statistics Education*, 1, 2007. `http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art5/`.

[61] V.Y. Kataoka, E.B. Ferreira, C.S.F. da Silva, and M.S. Oliveria. Increasing secondary students' statistical knowledge by focusing on teachers' engagement at Lavras, Minas Gerais, Brazil. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[62] G.E. Kennedy and Q.I. Cutts. The association between students' use of an electronic voting system and their learning outcomes. *Journal of Computer Assisted Learning*, 21:260–268, 2005.

[63] K.M. Kieffer, R.J. Reese, and B. Thompson. Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, 69:280–309, 2001.

[64] P.C. Knox. Using technology to support interactivity and improve learning in large group teaching. In *Thirty-second Annual Meeting of the Society for Neuroscience*, Orlando, FL, Nov 2002.

[65] C. Konold. Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, 3, 1995. `http://www.amstat.org/publications/jse/v3n1/konold.html`.

[66] D. Lass, B. Morzuch, and R. Rogers. Teaching with technology to engage students and enhance learning, 2007.

[67] R. Latessa and D. Mouw. Use of an audience response system to augment interactive learning. *Family Medicine*, 37:12–14, 2005.

[68] T.R. Lee. A study on the multimedia courseware for introductory statistics. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics*. International Statistical Institute, 1998.

[69] K. Lipson. Using the TI-83 graphics calculator in a liberal arts statistics course. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics*. International Statistical Institute, 1998.

[70] S.R. Luchini, M.P. Perelli D'Argenzio, and G. Moncecchi. Concept mapping for the teaching of statistics in primary schools: Results of a class experiment in Italy. In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics*. International Statistical Institute, 2002.

[71] E.L. MacGeorge, S.R. Homan, J.B. Dunning Jr., D. Elmore, G.D. Bodie, E. Evans, S. Khichadia, S.M. Lichti, B. Feng, and B. Geddes. Student evaluation of audience response technology in large lecture classes. *Educational Technology Research and Development*, 56:125–145, 2008.

[72] Z. Mahmud and C. Robertson. Developing and testing a teaching model using experimental design and interview analysis. In L. Pereira-Mendoza, editor, *Proceedings of the Fifth International Conference on Teaching Statistics*. International Statistical Institute, 1998.

[73] M. Martyn. Clickers in the classroom: An active learning approach. *EDUCAUSE Quarterly*, 30:71–74, 2007.

[74] E. Mazur. *Peer Instruction: A user's manual*. Prentice Hall, Upper Saddle River, NJ, 1997.

[75] I. McLeod, Y. Zhang, and H. Yu. Multiple-choice randomization. *Journal of Statistics Education*, 11, 2003. `www.amstat.org/publications/jse/v11n1/mcleod.html`.

[76] O. Meyer and M. Lovett. Implementing a computerized tutor in a statistical reasoning course: Getting the big picture. In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics*. International Statistical Institute, 2002.

[77] O. Meyer and C. Thille. Developing statistical literacy across social, economic and geographical barriers using a "stand-alone" online course. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[78] R.G. Miller, B.H. Ashar, and K.J. Getz. Evaluation of an audience response system for the continuing education of health professionals. *Journal of Continuing Education of Health Professionals*, 23:109–115, 2003.

[79] T.B. Miller and M. Kane. The precision of change scores under absolute and relative interpretations. *Applied Measurement in Education*, 14:307–327, 2001.

[80] G.A. Milliken and D.E. Johnson. *Analysis of Messy data, Vol 3: Analysis of covariance.* Chapman and Hall/CRC, New York, 2001.

[81] B. Morling, M McAuliffe, L. Cohen, and T.M. DiLorenzo. Efficacy of personal response systems ("clickers") in large introductory psychology courses. *Teaching of Psychology*, 35:45–50, 2008.

[82] V.N. Nair. Taguchi's parameter design: A panel discussion. *Technomerics*, 34:127–161, 1992.

[83] K.R. Nelms. Impact of hypermedia instructional materials on study self-regulation in college students. In *Annual Proceedings of Selected Research and Development [and] Practice Papers Presented at the National Convention of the Association for Educational Communications and Technology*, volume 1–2, Atlanta, GA, 2001.

[84] G. Norman. Rct = results confounded and trivial: The perils of grand educational experiments. *Medical Educator*, 37(7):582–584, 2003.

[85] T. Nosek, W. Wang, I. Medvedev, M. While, and T. O'Brian. Use of a computerized audience response system in medical student teaching: Its effect on active learning and exam performance. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education*, pages 2245–2250, San Diego, CA, 2006.

[86] J.C. Nunnally. *Psychometric Theory.* McGraw-Hill, New York, 1978.

[87] W.R. Penuel, C.K. Boscardin, K. Masyn, and V.M. Crawford. Teaching with student response systems in elementary and secondary education settings: A survey study. *Educational Technology Research and Development*, 55:315–346, 2007.

[88] J. Periasamy. Enhancing the link between statistics and mineral processing through project based learning. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics.* International Statistical Institute, 2006.

[89] M.S. Phadke. *Quality Engineering Using Robust Design.* Prentice-Hall, Englewood Cliffs, NJ, 1989.

[90] A. Pradhan, D. Sparano, and C.V. Ananth. The influence of an audience response system on knowledge retention: An application to resident education. *American Journal of Obstetrics and Gynecology*, 193:1827–1830, 2005.

[91] R.W. Preszler, A. Dawe, C.B. Shuster, and M. Shuster. Assessment of the effects of student response systems on student learning and attitudes over a broad range of biology courses. *CBE Life Sciences Education*, 6:9–20, 2007.

[92] H.C. Purchase, C. Mitchell, and L. Ounis. Gauging students' understanding through interactive lectures. In *Key Technologies for Data Management*, volume 3112/2004, pages 234–243. Springer, Berlin, 2004.

[93] S.W. Raudenbush and A.S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks, CA, 2002.

[94] R. Rogers. Using personal response systems to engage students and enhance learning, Nov 2003. Presented at Making Statistics More Effective in Schools and Business (MSMESB) Conference.

[95] R.J. Roselli and S.P. Brophy. Experiences with formative assessment in engineering classrooms. *Journal of Engineering Education*, 95:325–333, 2006.

[96] T.E. Schackow, M. Chavez, L. Loya, , and M. Friedman. Audience response system: Effect on learning in family medicine residents. *Family Medicine*, 36:496–504, 2004.

[97] C. Schau. Students attitudes: The "other" important outcome in statistics education. Alexandria, VA, 2003. Papers presented at the American Statistical Association Joints Statistical Meetings.

[98] C. Schau, J. Stevens, T.L. Dauphinee, and A. Del Vecchio. The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement*, 55:868–875, 1995.

[99] J.M. Shaughnessy. Research on statistics learning and reasoning. 2007.

[100] K. Siau, H. Sheng, and F.F.H. Nah. Use of a classroom response system to enhance classroom interactivity. *IEEE Transactions on Education*, 49:398–403, 2006.

[101] D. Stangl, D. Banks, L. House, and J. Reiter. Progressive mastery testing: Does it increase learning and retention? Yes and no. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics*. International Statistical Institute, 2006.

[102] P.S. Stein, S.D. Challman, and J.K. Brueckner. Using audience response technology for pretest reviews in an undergraduate nursing course. *Journal of Nursing Education*, 45:469–473, 2006.

[103] W.R. Stephenson. Statistics at a distance. *Journal of Statistics Education*, 9, 2001. `www.amstat.org/publications/jse/v9n3/stephenson.html`.

[104] J.R. Stowell and J.M. Nelson. Benefits of electronic audience response systems on student participation, learning and emotion. *Teaching of Psychology*, 34:253–258, 2007.

[105] M.L.M.M. Sundefeld, A.V. Guimaraes, C. dos Anjos Santos, G.K. Kavano, R.Y. Takano, A.B.S.P. Fernandes, and I.C.L. Poi. An interactive CD-Rom to teach statistics applied to health notions to students in the fundamental schooling. In A. Rossman and B. Chance, editors, *Proceedings of the Seventh International Conference on Teaching Statistics.* International Statistical Institute, 2006.

[106] G. Taguchi. *Introduction to Quality Engineering.* Asian Productivity Organization, Tokyo, 1986.

[107] P.J. Trapskin, K.M. Smith, J.A. Armitstead, and G.A. Davis. Use of an audience response system to introduce an anticoagulation guide to physicians, pharmacists, and pharmacy students. *American Journal of Pharmaceutical Education*, 69:Article No. 28, 2005.

[108] M. Uhari, M. Renko, and S. Hannu. Experiences of using an interactive audience response system in lectures. *BMC Medical Education*, 3, 2003.

[109] J. Utts, B. Sommer, C. Acredolo, M.W. Maher, and H.R. Matthews. A study comparing traditional and hybrid internet-based instruction in introductory statistics classes. *Journal of Statistics Education*, 11, 2003. `www.amstat.org/publications/jse/v11n3/utts.html`.

[110] L.A. Van Dijk, G.C. Van Den Berg, and H. Van Keulen. Interactive lectures in engineering education. *European Journal of Engineering Education*, 26:15–28, 2001.

[111] B. Ward. The best of both worlds: A hybrid statistics course. *Journal of Statistics Education*, 12, 2004. `www.amstat.org/publications/jse/v12n3/ward.html`.

[112] J.M. Watson and B.A. Kelly. Can Grade 3 students learn about variation? In B. Philips, editor, *Proceedings of the Sixth International Conference on Teaching Statistics.* International Statistical Institute, 2002.

[113] L. Wilkinson and APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54:594–604, 1999.

[114] R.H. Williams and D.W. Zimmerman. Are simple gain scores obsolete? *Applied Psychological Measurement*, 20:59–69, 1996.

[115] E. Wit. Who wants to be... The use of a personal response system in statistics teaching. *MSOR Connections*, 3:14–20, 2003.

[116] C.F.J. Wu and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization.* John Wiley and Sons, New York, 2000.

[117] E. Zhu. Teaching with clickers, 2007.

[118] I.A. Zualkernan. Using Soloman-Felder learning style index to evaluate pedagogical resources for introductory programming classes. In *Proceedings of the Twenty-ninth International Conference on Software Engineering*, pages 723–726, Minneapolis, MN, May 2007.