UCLA
Technology Innovations in Statistics Education

Title
A Randomized Experiment Exploring How Certain Features of Clicker Use Effect Undergraduate Students' Engagement and Learning in Statistics

Permalink
https://escholarship.org/uc/item/2503w2np

Journal
Technology Innovations in Statistics Education, 4(1)

ISSN
1933-4214

Authors
McGowan, Herle M
Gunderson, Brenda K

Publication Date
2010-06-04

DOI
10.5070/T541000042

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

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. James 2006; Duncan 2005; Fagan, Crouch, and Mazur 2002). They are also gaining attention in other fields, such as medicine (e.g. Nosek, Wang, Medvedev, While, and O'Brian 2006; Pradhan, Sparano, and Ananth 2005; Schackow, Chavez, Loya, and Friedman 2004), engineering (e.g. Zualkernan 2007; Demetry 2005; Siau, Sheng, and Nah 2006), biology and life sciences (e.g. Freeman et al. 2007; Preszler, Dawe, Shuster, and Shuster 2007; Brewer 2004), psychology (Cleary 2008; Morling, McAuliffe, and Cohen 2008), accounting (Beekes 2006; Carnaghan and Webb 2006), agriculture (Conoley, Moore, Croom, and Flowers 2006), computer science (Kennedy and Cutts 2005), earth science (Greer and Heaney 2004), and statistics (Rogers 2003; Wit 2003). More recently, clickers have been incorporated into elementary and secondary education (Chen et al. 2005; Conoley et al. 2006; Penuel, Boscardin, Masyn, and Crawford 2007; Hanley and Jackson 2006). Several good overviews of their uses, including guides for writing good conceptual clicker questions exist in the literature, such as Beatty (2004), Duncan (2005), Beatty, Gerace, Leonard, and Dufresne (2006), and Zhu (2007).

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 (e.g. Caldwell 2007; Demetry 2005; Roselli and Brophy 2006). 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 (Caldwell 2007; Jackson and Trees 2003; Roselli and Brophy 2006). Also, the interactive nature of clickers may help students pay more attention to each question (Latessa and Mouw 2005; Miller, Ashar, and Getz 2003; Uhari, Renko, and Hannu 2003). Many students who have used clickers report that they improve the classroom experience (e.g. Auras and Bix 2007; MacGeorge, Homan, Dunning Jr. et al. 2008; Trapskin, Smith, Armitstead, and Davis 2005) and improve their own understanding of the material taught (e.g. Preszler et al. 2007; Bunce, VandenPlas, and Havanki 2006; Trapskin et al. 2005).

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 (2006) 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. Van Dijk, Van Den Berg, and Van Keulen (2001) 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 (2007) 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 may be less comfortable asking them.

In terms of learning, many studies have found higher exam scores when clickers were used (e.g. Conoley et al. 2006; Freeman et al. 2007; Pradhan et al. 2005). It should be noted, though, that several only demonstrated conditional improvement. For example, Carnaghan and Webb (2006) and Schackow et al. (2004) found a significant improvement in scores only for those exam questions that were most closely related to the clicker questions asked during class. Several authors (Kennedy and Cutts 2005; Lass, Morzuch, and Rogers 2007; Nelson and Hauck 2009; Nosek et al. 2006) 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. Two studies addressed this concern. One looked formally at student ability, as measured by performance on a prerequisite course's exam, and found that those students who were previously poor performers (scored less than 60%) actually gained more from the use of clickers than other students. They also looked at rate of clicker use and found that those who used clickers more (answering more than 52% of the clicker questions) preformed significantly better than those who used clickers less, despite there being no significant difference in average scores on the prerequisite course exam. Another study formally manipulated the number of clicker questions asked during a semester: Preszler et al. (2007) 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.

Several studies found no significant difference in exam scores for students who used clickers versus those who did not (e.g. Miller et al. 2003; Schackow et al. 2004; Dill 2008). One study even found significantly worse exam scores for students using clickers: Van Dijk et al. (2001) 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 1997 for more on Peer Instruction). They found that students in the clicker-only group had lower exam scores than students in the other two groups, which were similar in performance to each other. Van Dijk et al. (2001) attributed this lower performance to the fact that students in the clicker-only group seemed to ask fewer clarifying questions.

While most studies of the effects of clicker use on learning looked only at short-term improvement, one study looked at retention of knowledge. Crossgrove and Curran (2008) found that non-majors taking an introductory biology course remembered more of the material that had been taught with clickers than without, as measured by exam performance four months after the course had been completed. They also looked at retention of material taught with clickers in an upper-level course for biology majors, but found no significant difference in exam scores at the end of the course and at the four-month follow-up.

1.1 Limitations of Previous Clicker Research

There have been two primary limitations of previous research on the effectiveness of clickers. First, there has been little connection to theory or existing studies. While most proponents of clicker use claim they improve student engagement and learning, few have provided reasoning rooted in cognitive psychology as to why. The most striking exception to this is Mayer et al. (2009), which included a thorough review of the literature on using questioning methods to improve student learning. In this study, three groups were compared: 1) students that used clickers to answer multiple choice questions during lecture, 2) students that answered multiple choice questions during lecture without clickers, using instead both a show of hands and written responses to the same questions, and 3) students that did not answer multiple choice questions during lecture. Using their literature review as a foundation, the authors created a model that explained why they thought asking questions during lecture would lead to improved learning:

> Thus, our main prediction is that the clicker treatment will lead to greater student-teacher interaction, which encourages deeper cognitive processing during learning, which in turn will be reflected in improvements in exam score in the course. In short, we expect the clicker group to produce higher exam scores than the control group. If we are successful in implementing the questioning method without computer-based technology in the no-clicker group, we also expect the no-clicker group to outperform the control on exam scores and to be equivalent to the clicker group. (p. 53)

Mayer et al. (2009) found that the clicker group had higher exam scores on average than either of the other two groups, which were statistically indistinguishable from each other. They attribute these results to the relative ease of collecting student responses with clickers, which was less disruptive than when written responses had to be physically collected by the instructor. While their learning theory focused on the pedagogical practice of questioning, this study also illustrates one of the advantages of the clicker technology itself, namely the ease of implementation.

The second limitation of clicker research is methodological confusion between the treatment of interest (roughly, "clicker use") and the simple pedagogical change of asking more interactive questions in class. Many studies, which compared classes that used clickers to classes that did not, failed to make this distinction. 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 (Carnaghan and Webb 2006). A few studies did address this design flaw. For example, Schackow et al. (2004) and Carnaghan and Webb (2006) used crossover designs where students responded to multiple-choice questions verbally or with clickers, and both found some evidence of increased performance when clickers were used. Freeman et al. (2007) 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. No significant difference in exam performance was found between the groups, though attendance was higher in the clicker sections (Freeman et al. 2007).

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–while addressing some of the methodological issues discussed here, an experiment was conducted from January to April 2008 at a large mid-western university. This experiment took place in the laboratory sections of a multi-section introductory data analysis course.

1.2 Description of the Course

The course in which the experiment was implemented was a 4-credit course taught every semester (14 week term) at the university. 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 included descriptive statistics (numerical and graphical summaries), probability, sampling distributions, and inference procedures. The inference procedures included 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. Students attended three hours of lecture and one 1.5 hour computer lab each week. The lecture sections varied in size, ranging from 60 students to over 400 students. The schedule of the lecture sections also varied: sections were 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 was covered in all lecture sections. During the experimental semester there were six lecture sections taught by a team of four instructors.

Lab sections were more uniform than lecture sections in terms of size and structure; there were also many more lab sections, which allowed for replication of treatment conditions. For these reasons, the experiment was implemented in the lab sections of the course. The goal of the labs was to reinforce concepts presented in lecture and provide hands-on examples of data analysis using the statistical analysis package SPSS. Occasionally, however, some material was 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—were 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.

2. DESIGN AND RESEARCH QUESTIONS

The research questions, outcomes, treatment variables were selected to formally test the oft-touted benefits of clickers, namely that they increase students' engagement and learning. As such, these two constructs were the primary outcomes of interest in this experiment. The terms “engagement” and “learning” 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 (e.g. Fredricks, Blumenfeld, and Paris 2004). 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; in this experiment, however, all three aspects of engagement were considered (see Section 3).

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). 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, since courses differ with respect to the topics emphasized and the exam structure. For example, one course exam could focus on computation while another course exam could focus on interpretation, so that the same score on these different exams would not necessarily imply the same knowledge of statistics. To avoid this problem, 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).

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. For example, clicker users tend to champion their strength for providing immediate feedback to both students and instructors, without systematically considering the amount or timing of this feedback. However, the experience of the first author in teaching with clickers seemed to indicate that there might be practical limits on how to provide this feedback. For logistical reasons, clickers were initially used in the course only a few times during a semester, for entire class periods. During these classes, students often became distracted or disruptive while waiting for others to enter their answer to a question. This experience seemed to indicate the possibility of an "overdose," so to speak, of clicker use—a possibility which had not been widely considered in the existing literature. To address this, two of the treatment variables in this experiment were the number of questions asked with clickers during a lab session and the placement of those questions throughout the material (specifically, if the questions were asked in a group or more spread out). A third treatment variable was considered to explore the effect of external incentives (e.g. grades) in getting students to use the clickers. As described previously, most studies on clickers contain student-reported data on their positive effects on the classroom environment or the attention paid during class. If these reported benefits are true, one would expect that students would choose to use clickers even when it is neither required nor tracked. In contrast, if these reported benefits are not true or not great in magnitude, students may not bother using clickers when it is not required of them. (As a side note, there are clearly many other features of clicker use that could have been explored, but these three were of particular interest to the authors for the reasons stated here. A discussion of possibilities for future research is provided in Section 7.) The three treatment variables were operationally defined as follows:

1. Frequency: The number of clicker questions asked during a lab session
    (a) *High*: At least 6 clicker questions were asked
    (b) *Low*: 3-4 clicker questions were asked
2. Agglomeration: Asking all questions consecutively in an "agglomeration" or group
    (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, an "agglomeration" 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 grades assigned

Two experimental designs—a factorial design and a crossover design—were used simultaneously to investigate the effect of these three components on both student engagement and learning. The factorial experiment was used to explore the effects of Frequency and Agglomeration, as well as their interaction with each other (where a negative interaction would represent an "overdose" of clicker use). Guiding the exploration of these two treatment variables were the following research questions:

RQ1. What is the main effect of Frequency?

1. It could be positive, if students value the instant feedback characteristic of clickers.
2. It could be negative, if students lose interest due to system overuse.

RQ2. What is the main effect of Agglomeration?

1. It could be positive, if grouping similar questions helps to reinforce a concept.
2. It could be negative, if students lose interest due to system overuse.

RQ3. Is there a negative interaction between Frequency and Agglomeration?

1. This could indicate that students lose focus when too many clicker questions are asked consecutively.

The crossover design was used to investigate the effect of External Incentive on behavioral engagement–namely, whether students choose to use the clickers when it was not required. All students were required to purchase a clicker for the course; therefore 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 (James 2006). For the *Moderate* level, the incentive of grading is removed, but there may still be some incentive from being monitored by the instructor. 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 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. These theories lead to the final research question to be explored in this experiment:

RQ4. Do students perceive a value to using clickers even when their use is neither required nor monitored?

If the answer to this question is "yes," that clickers do engage students, then the rate of clicker usage should be similar across all levels of External Incentive. If the answer is "no," then the rate of clicker usage should decrease as the level of External Incentive decreases.

Any student who was registered for course after the university drop/add deadline was eligible to participate in this study. There were a total of 1277 students enrolled in the class during the experimental semester, 1197 (94%) of which consented to allow their data to be used in analyses (see Section 4 for more on consent and implementation). Students in the course were divided between 50 lab sections taught by 24 GSIs. 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 1.

Table 1: Design of Factorial Experiment

		Agglomeration	
		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 corresponding participation rate for that group.

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, are presented in Table 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 2: Design of Crossover Experiment

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 that consented to have their data used. The number in parentheses is the participation rate for that sequence.

3. MEASURES

Three aspects of engagement—emotional, cognitive, and behavioral—were considered in this experiment. Emotional and cognitive engagement were measured through student report of attitudes towards statistics and clickers, using several subscales (Affect, Value, Emotion, and Cognitive Competence) of the Survey of Attitudes Towards Statistics (SATS, Schau, Stevens, Dauphinee, and Del Vecchio 1995) as well as questions developed by the Center for Research on Learning and Teaching at the University of Michigan. Behavioral engagement was measured by the percent of students per lab section that used clickers under each level of External Incentive, where two levels of “clicker use” were considered, to account for the varying number of clicker questions asked: 1) Answering at least one clicker question, or 2) Answering at least 50% of the clicker questions 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* level.

Learning was measured using several instruments from the ARTIST project, including the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS; delMas, Garfield, Chance, and Ooms 2006) and four topic scales (Normal Distribution, Sampling Distributions, Confidence Intervals, Significance Tests). The ARTIST topic scales served as proximal measures of learning, since 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. In contrast, CAOS is a comprehensive exam which served as the measure of both pretreatment knowledge of statistics and longer-term learning. Each of the outcome measures was selected for use in this experiment because they are nationally available and have 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 Agglomeration, and -1 for both the *Low* level of Frequency and the *On* level of Agglomeration. Additionally, the actual number of clicker questions asked during each lab was reported by the GSI. 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 the current statistics course)
- Work: Typical 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 the course prior to the start of the experimental semester

Each covariate was examined for imbalances between the treatment groups, some of which were found. These imbalances 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 were with the covariates Year and GSI Experience: The Blue Team had a disproportionately large number of Freshman and small number of Juniors and Seniors; additionally, the Yellow Team had a disproportionately larger average GSI experience. To account for any pretreatment discrepancies between groups, covariate selection was used to identify important variables for inclusion in each of the regression models considered.

4. IMPLEMENTATION

The treatment period did not begin until after the university 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.
- 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, at which time students were given a formal description and asked to provide or refuse their consent to have their data used for analysis. 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 participated in experimental procedures—students provided consent only to allow their data to be analyzed—which was important for two reasons. First, all students were treated in the same manner regardless of their desire for their data to be used, simplifying planning for the GSIs and the researchers. Second, it was expected to lead to a higher consent rate since students were not required to do any additional work (recall that the consent rate was 94%, so this approach seemed to be successful).

After the consent process in the third week of labs, 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 this week, students completed the ARTIST topic scale about the normal distribution. The other three topic scales were completed approximately every other week after that. Post treatment administration of CAOS and the attitudes survey took place during the final week of labs.

Throughout the treatment period, several clicker questions were asked in each lab. Clicker questions were selected from the lab workbook for the course, which primarily contained activities to help students apply concepts and procedures that had been learned during lecture. In fact, all lab sections were asked the same questions (with the same answer choices, where appropriate) each week. This was done to avoid a previously discussed limitation of many clicker studies, where there was confusion between clicker use and the pedagogical technique of asking questions. From the total set of questions for a given week, the first author selected particular questions to be asked with clickers for each lab section. The sections thus 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 grouped together) and the level of external incentive in encouraging students to use the clicker remotes.

At staff meetings during each week of the treatment period, the first author provided the set of questions to be asked during the following week, indicated which questions should be asked with clickers for each GSI, and also reminded each GSI of their team for the factorial experiment and the appropriate level of External Incentive they should be running for the crossover experiment. While GSIs were told whether or not clicker questions should be asked in an agglomeration for their lab, they were not given much guidance on the specific placement of questions within their lessons. Likewise, GSIs were allowed to decide how to follow-up on each question asked with clickers—if and how they would discuss the question with their students. This allowed each GSI to better plan their lesson in accordance with their teaching style.

At the end of each class period, GSIs were asked to complete a brief survey on how well the planned implementation procedures were followed. GSIs reported the number of questions asked with clickers; if fewer than the intended number of clicker questions were asked, GSIs also reported the reason why (e.g. technology problems; running out of time). In addition, GSIs reported what level of External Incentive had been run, as well as whether or not an announcement of this level was made to the class (as this factor could not affect student behavior if students did not know what the appropriate level was for that week). A discussion of implementation infidelity and the implications it may have had on the results of this experiment is provided in Section 6.5.

5. ANALYSIS OF THE EXPERIMENT

This section presents analyses of all outcomes pertaining to engagement and 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. Discussion of the results is presented in Section 6.

5.1 Emotional and Cognitive Engagement

Recall that statements on the attitude survey were drawn from the SATS 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 and are the only statements specific to the technology used in this experiment. 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 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 4 presents the same information for the post treatment average ratings.

Cronbach's α (Nunnally 1978) 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 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. 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 (Schau 2003).

Figure 1 plots the average of the mean post treatment ratings by treatment factor for each subscale on the attitude survey. In each panel, there appears to be an interaction, though the magnitudes of the differences between the team averages are small. To test the significance of any effects, each of these scales was used as the response in a hierarchical linear model (HLM) that included nested random effects for GSI and lab section. 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. The effect of treatment on each response was estimated with terms for the main effects of and interaction between Frequency and Agglomeration. For each model, a 5% significance level was used to determine statistical significance for the main effects, while a 10% level was used for the effect of the interaction. Unfortunately, for each of the five models, there were no significant effects of treatment after adjusting for potential confounding variables (*p*-values > 0.1 in all cases; models not shown).

In analyses discussed so far, 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 (models not shown). Seven statements showed significant effects of the design factors:

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

Table 3: 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

Table 4: 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*, Agglomeration=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

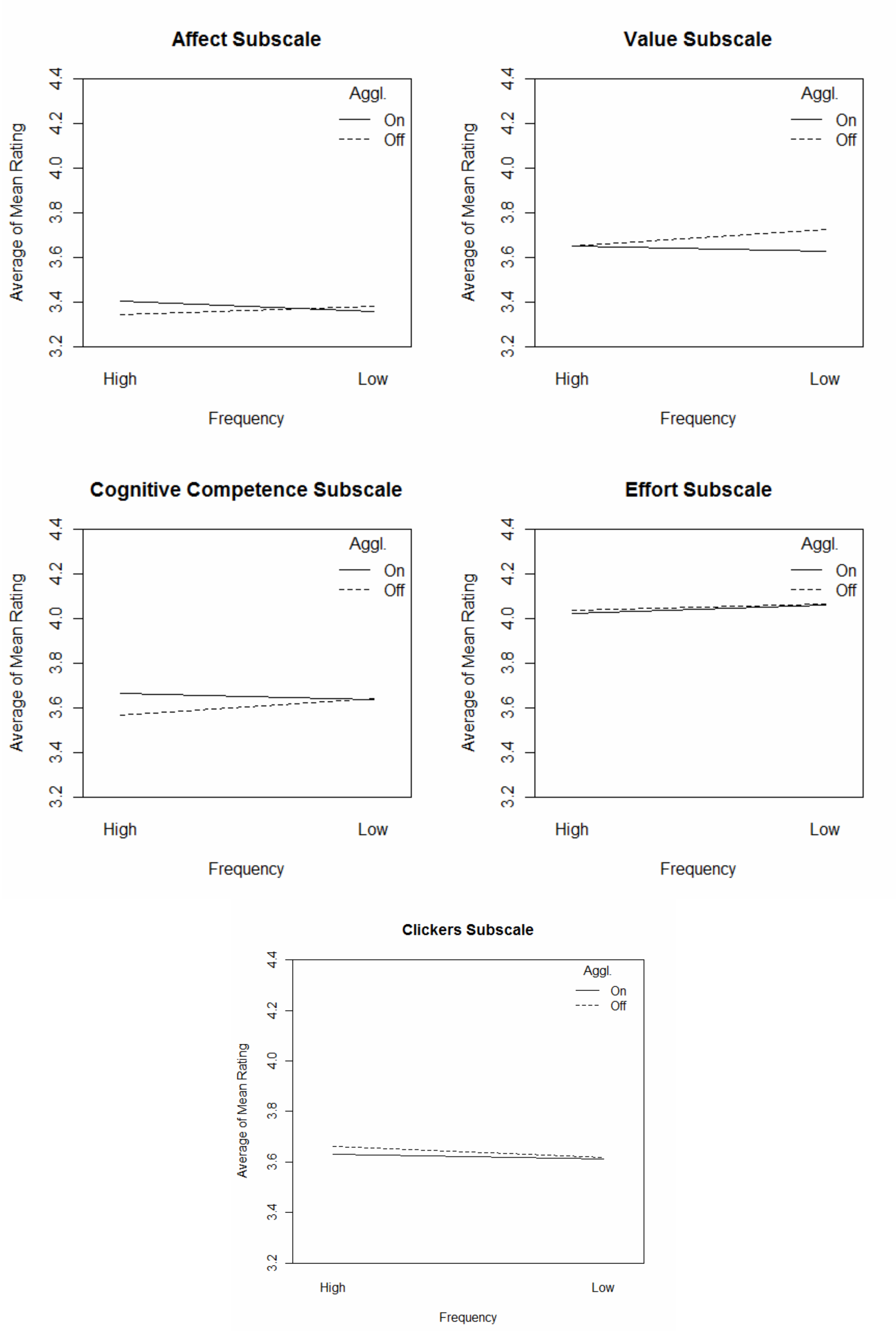

Figure 1: Average Mean Post Treatment Ratings by Design Factor for each Attitude Subscale
In each panel: the y-axis is scaled to have the same range; the solid line corresponds to Agglomeration *On* and the dashed line to Agglomeration *Off*.

Several consistent patterns could be seen in the models for these seven statements. First, when the effect of Frequency was significant at the 5% level, it was positive, indicating that asking more clicker questions is better. When the interaction between Frequency and Agglomeration was significant at the 10% level, it was negative. Interestingly, the effect of Agglomeration (when significant at the 5% level) was 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. Considering the statement pertaining specifically to clickers, 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. Another consistent pattern was that the largest probability of moving from a pretreatment rating of "Neutral" to a post treatment rating of "Agree" generally occurred for the Yellow Team (Frequency = *High*, Agglomeration = *Off*). For all teams and across all statements, the probability of making this improved rating was encouragingly high, ranging from 28% to 56% and often higher than making the change to a negative rating of "Disagree." An exception to this was the last statement, "I use statistics in my everyday life." Finally, the probabilities of improving from a pretreatment rating of "Disagree" to a post treatment rating of "Agree" ranged from 12% to 36% across all teams and statements (excluding the last statement). While these probabilities were understandably lower than those for moving from "Neutral" to "Agree", they were still encouraging.

5.2 Behavioral Engagement

Recall that 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 (see Table 2).

Two analyses of clicker use were performed, to reflect the discrepancy in number of clicker questions asked at the *High* and *Low* levels of Frequency. For the first, clicker use was defined as 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. The pattern of results was nearly identical for each of these analyses, with the overall proportion of users being slightly lower under the stricter criterion (answering at least 50% of the clicker questions). For sake of space, only the results for this criterion will be presented in detail here.

Figure 2 shows the proportion of students in each sequence who answered at least 50% of the clicker questions for a particular week of the semester. All sequences show some decrease in the proportion of users over the course of the treatment period, with sequences 3 and 4—in which the level of External Incentive declines over the semester—showing the largest declines. For this data, a hierarchical linear model was fit that included random effects for GSI and that was weighted by the number of students in attendance for a particular lab section and week. Here the response was the number of students in each lab section answering at least 50% of the clicker questions for a given week. Table 5 shows the final results for this model. 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.

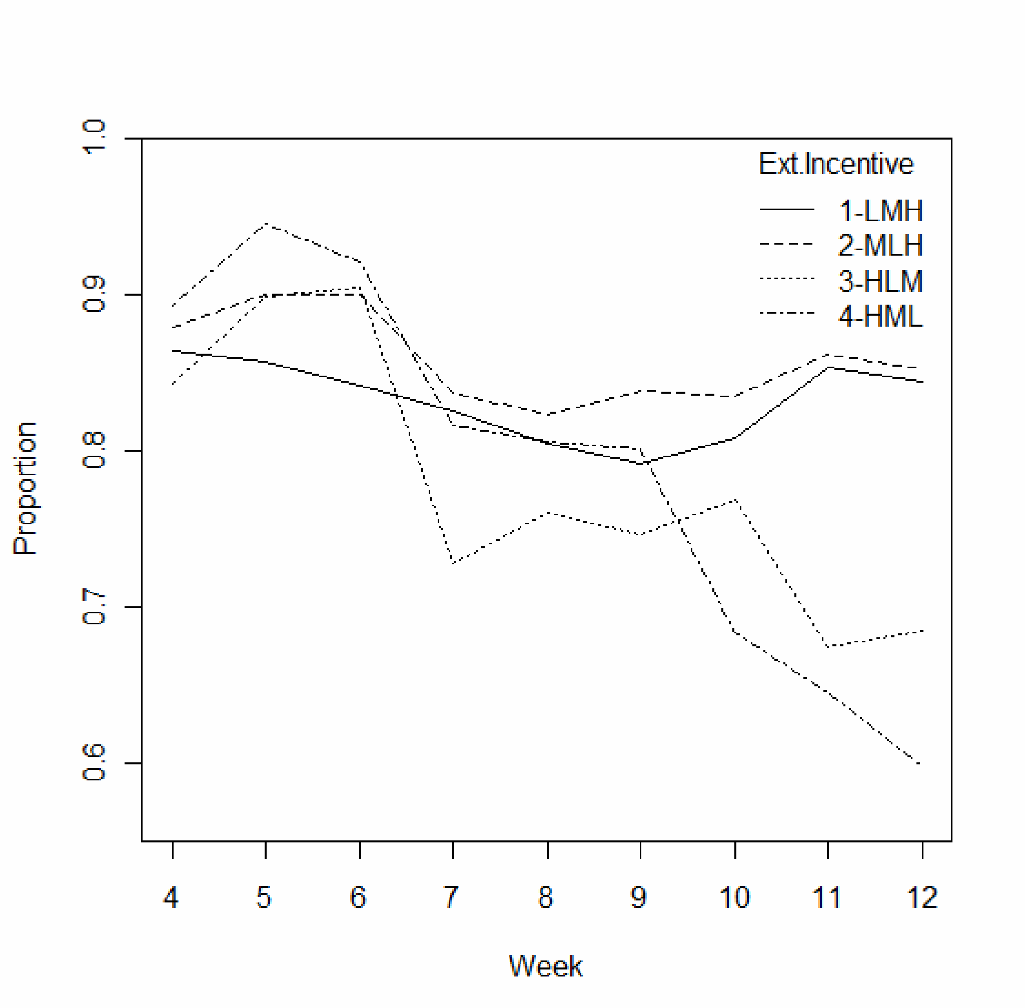

Figure 2: Proportion of Students Answering At Least 50% of the Clicker Questions
The solid represents the proportion, for each week of the treatment period, of students in sequence 1 (*Low-Mod-High* External Incentive) who answered at least 50% of the clicker questions; the dashed line represents the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

Table 5: HLM Results for Behavioral Engagement
Number Answering At Least 50% of 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

5.3. Learning: The Comprehensive Assessment of Outcomes in a First Course in Statistics

The primary measure of learning for this experiment was the CAOS instrument. Students first completed CAOS after the drop/add deadline had passed, when course enrollment was fixed (with the exception of a handful of students who dropped late). 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 correctly answered about 52% of the questions, on average, at this time. Students also completed CAOS during the last lab session of the semester. Descriptive statistics for CAOS, for the entire sample (Overall) and by treatment group (Team), are given in Table 6. 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 by the post treatment exam. The treatment groups had roughly equivalent scores on the pretreatment CAOS, with the Green Team (Frequency = *Low*, Agglomeration = *Off*) having a slightly higher mean than the other teams. Overall, the average CAOS score increased by 13.7% (equivalent to five and a half points) from pre to post treatment.

Table 6: Descriptive Statistics for CAOS

	Team[a]	Cronbach's α	Min	Median	Mean (SD)	Max	N
Pretreatment	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
Post treatment	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	0.74	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*, Agglomeration=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

Figure 3 plots the average percent correct on the post treatment CAOS by treatment factor. Interestingly, the lines in this picture appear parallel, indicating that there is no interaction between Frequency and Agglomeration. However, this plot does not account for possible confounding factors. To test if Frequency and Agglomeration had significant effects while accounting for possible confounders, a hierarchical linear model was fit including nested random effects for GSI and lab. The response was the percent correct on the post treatment CAOS. Table 7 shows the final results for this model. After adjusting for several important confounders, the main effect of Frequency is estimated to be -1.370 percent; the main effect of Agglomeration 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 one point (out of 40 points possible) on CAOS. While neither of the main effects are significant at the 5% level, the interaction is significant at the 10% level. This analysis indicates that, holding all else equal, asking a low number of clicker questions throughout a class led to an increase of 4.469 percent correct, or roughly two points, on the post treatment CAOS (as compared to asking a high number of clicker questions consecutively).

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. Covariate selection was used with each of the resulting three-quarter subsamples and the final validation models were examined for consistency with the model presented in Table 7. The overall substantive conclusions about the magnitude and significance of the design factors were consistent for each validation model (not shown).

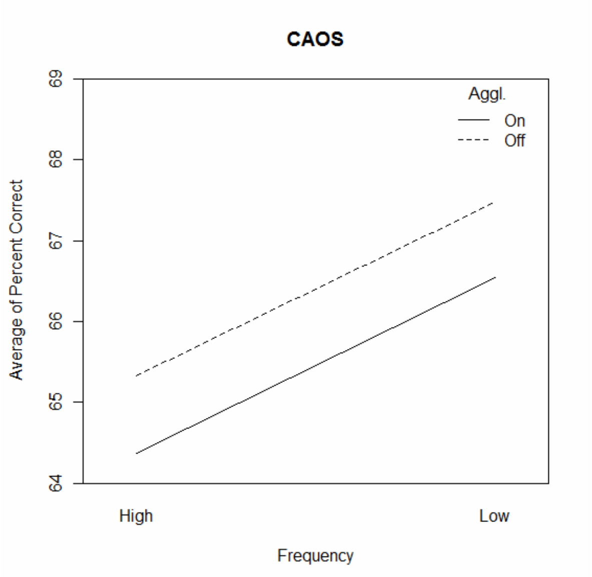

Figure 3: Average Percent Correct for Post Treatment CAOS by Treatment Group
The solid line corresponds to Agglomeration *On*, the dashed line to Agglomeration *Off*.

Table 7: 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
Agglomeration	1.605	0.448	17	0.091
Interaction	-1.494	0.413	17	0.088

Note: Estimates reported for Frequency, Agglomeration, 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.

Since the 40 CAOS questions were not of equal difficulty, several descriptive analyses were undertaken to explore the performance of the treatment groups (Team) by question. Figure 4 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. For the most part, each team shows improvement from pre to post treatment and often the Blue Team (Frequency=*Low*, Agglomeration=*Off*) performs the best. While linear regressions lines are not the best fit for this data, they do provide an idea of the average performance for each team. The line that stands out the most is that of the Blue Team, indicating that asking a few clicker questions throughout a class results in the highest percentage of correct responses, on average. Similar results were seen when questions were grouped by statistical concept and average performance for each concept was considered (results not shown). Each team typically showed improvement from pre to post treatment, and there were a few topics for which the Blue Team performed best. These topics included: confidence intervals (CAOS questions 28-31), making sense of data (11-13, 18), understanding distribution (1, 3-5), reading a histogram (6, 33), and gathering data (7, 38, 22, 24).

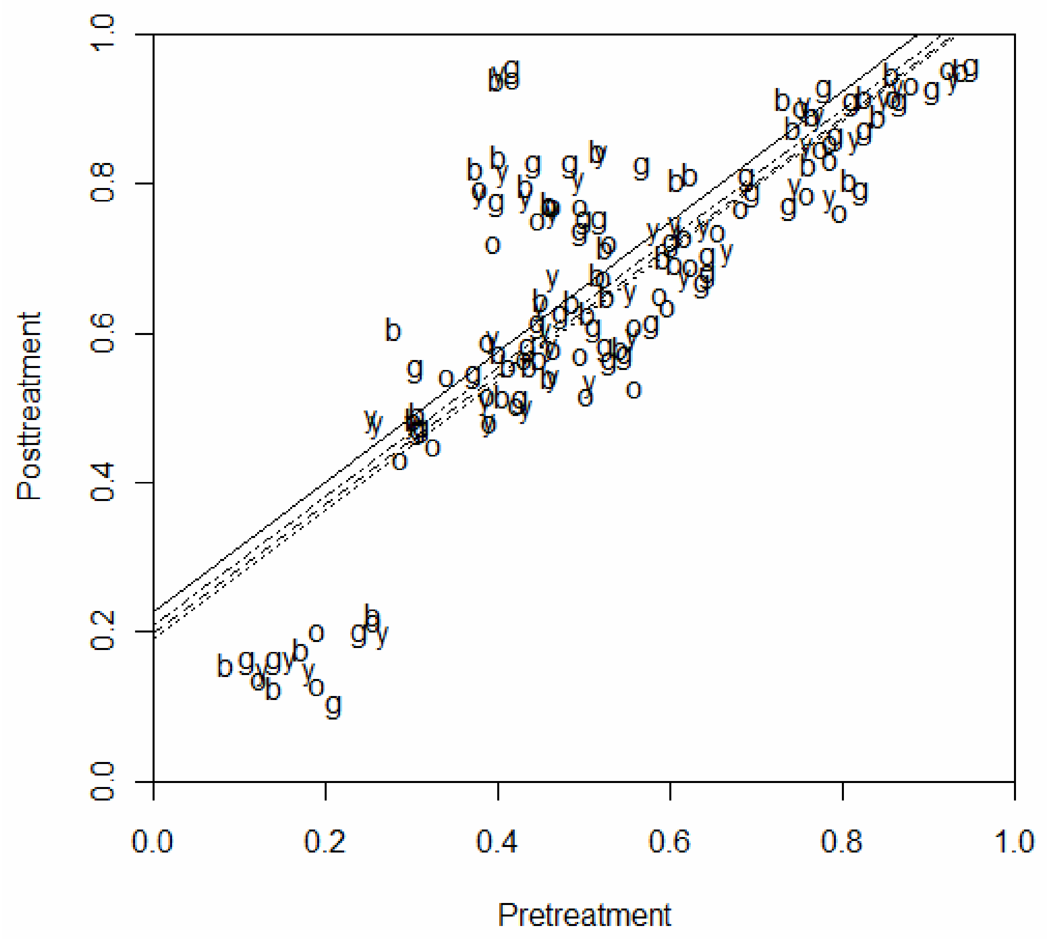

Figure 4: Proportion of Correct Responses for Each CAOS Question by Team.

Plotting character corresponds to team name: g=Green (Frequency=*Low*, Agglomeration=*On*); b=Blue (*Low*, *Off*); *Off* =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.

5.4 Learning Outcome: ARTIST Topic Scales

Table 8 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*, Agglomeration = *Off*) had the highest average score for each topic scale.

Figure 5 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 *Off* level of Agglomeration 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, Agglomeration, or their interaction (results not shown). Given the extremely low reliabilities shown in Table 8, this is not surprising. Because of this, further analysis of the topic scale data was not conducted.

Table 8: 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*, Agglomeration=*On*); Blue (*Low*, *Off*); Orange (*High*, *On*); Yellow (*High*, *Off*).

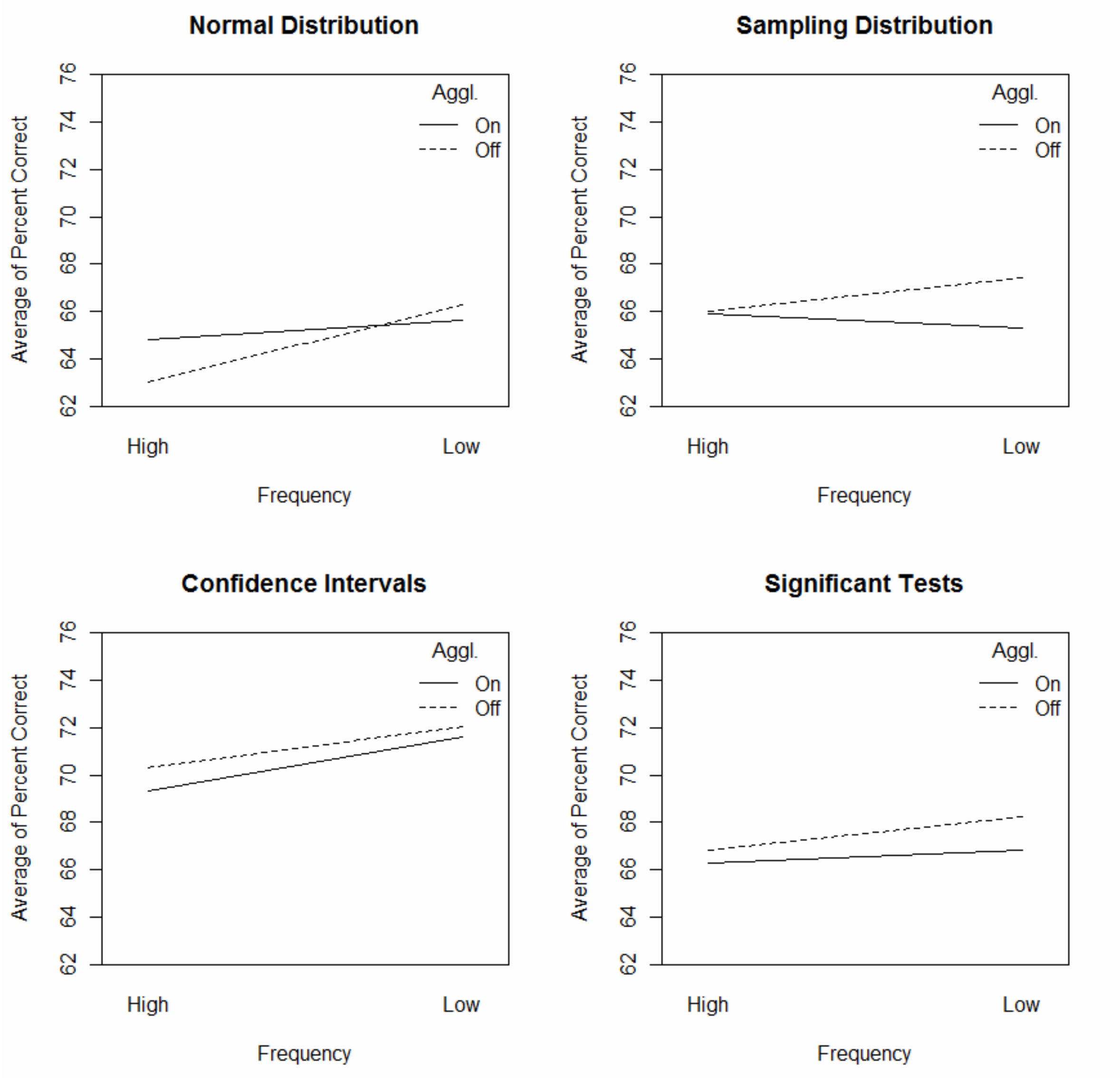

Figure 5: Average Percent Correct for the ARTIST Topic Scales
In each panel: the y-axis is scaled to have the same range; the solid line corresponds to Agglomeration *On* and the dashed line to Agglomeration *Off*.

6. DISCUSSION: THE EFFECT OF CLICKERS ON ENGAGEMENT AND LEARNING

In Section 2, the relevant research questions for the factorial experiment were presented as:

RQ1. What is the main effect of Frequency?
RQ2. What is the main effect of Agglomeration?
RQ3. Is there a negative interaction between Frequency and Agglomeration?

The relevant research question for the crossover experiment was:

RQ4. Do students perceive a value to using clickers even when their use is neither required nor monitored?

Discussion about each research question follows.

6.1 Discussion of RQ1

Table 9 shows the estimated main effects and standard errors for Frequency, Agglomeration, and their interaction from the hierarchical analyses of engagement and learning outcomes. As can be seen from the first column, the main effect of Frequency on engagement was estimated to be positive for each attitudinal outcome—indicating that asking more than 6 clicker questions is better than asking 3-4 questions—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 was 1.4 percent (0.56 points on the 40-point scale for CAOS) and was not significant at the 5% level.

Table 9: Summary of Effect s of Design Factors on Learning and Engagement

	Frequency		Agglomeration		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 are several possible explanations for these results, the simplest being that 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 these treatment variables. For example, clicker questions were based on existing questions in the lab workbook. This was done to ensure seamless incorporation of clicker questions within the already busy labs, but the activities and questions included in this workbook tend to be 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–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 or recall questions (i.e. Do you remember that definition?).

6.2 Discussion of RQ2

From the second column of Table 9, the main effect of Agglomeration 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 Agglomeration on learning was estimated to be positive–indicating that incorporating clicker questions throughout a class is better than asking them consecutively. The effects of Agglomeration on learning were larger than the effects of Agglomeration 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 Agglomeration = *Off*), tended to outperform the teams where Agglomeration = *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 grouped 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 question asked after discussion of the topic could require students simply to remember what they had been told (Beatty 2004).

6.3 Discussion of RQ3

Looking at the final column of Table 9, the effect of the interaction between Frequency and Agglomeration was estimated to be positive 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 for the existence of a negative interaction, indicating that that asking too many clicker questions consecutively is not conducive to engagement or 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 Agglomeration.

6.4 Discussion of RQ4

Table 10 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, 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 6 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 Figure 2).

Table 10: Summary of Effects of External Incentive on Behavioral Engagement

	External Incentive			
	Moderate		*High*	
Clicker Use	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

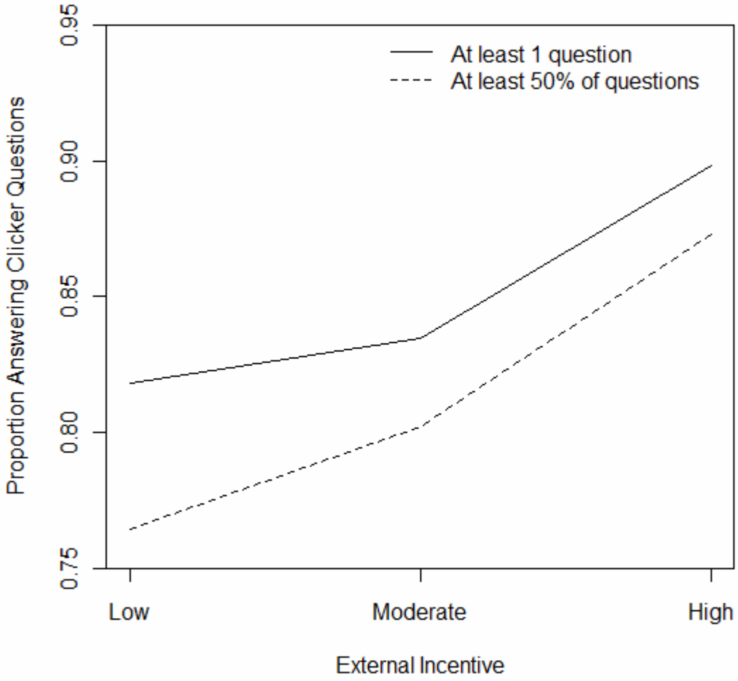

Figure 6: Proportion of Students Using Clickers by Level of External Incentive

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.

6.5 Discussion of Implementation Procedures

As discussed in Section 4, restrictions on GSIs were kept to a minimum so that they could teach with their own style. This was done primarily to avoid conflicts in the team or with the experimental procedure. In hindsight, however, the guidance provided was not enough, especially with respect to the placement of clicker questions. 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 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, and possibly the overall impact of the clickers.

In addition, 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. In about 4% of all lab sections over the nine weeks of the treatment period, no questions could be asked with clickers due to technical or other issues. This would be especially problematic when it occurred in those labs assigned to the *High* level, as they were essentially running at *Low* Frequency for those sessions (which account for about half of the instances where no clicker questions could be asked). It is possible that this could have affected the ability to assess the effect of Frequency. 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 a memo provided during the weekly staff meetings, 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.

If this experiment were to be run again, we would improve the implementation procedures by better emphasizing the switch between the various levels of External Incentive throughout the semester, as well as importance of announcing the level to students each week. Additionally, we would provide more detailed lesson plans to each GSI so that the placement of clicker questions was more tightly controlled. Realistically, it could be difficult to provide lessons that specify the placement of each clicker question for such a large experiment over such a long period of time, and even providing such plans would not guarantee GSI compliance with them. It might be possible then to improve implementation by shortening the length of the treatment period to just a few weeks instead of nine.

7. BRIEF SUMMARY, CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

This paper summarized the design and analysis of an experiment on the use of clickers in an introductory statistics course. The experiment had two main designs, run concurrently. First, 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 (Agglomeration) had on emotional and cognitive engagement as well as on learning. Second, 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.

Hierarchical linear models including nested random effects for GSI, lab and student were fit for several outcomes. 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.

The discussions in Sections 6.1 and 6.2 point to the importance of having a sound pedagogical purpose for each clicker question. For example, a question could be designed to elicit deeper thought about a concept, or to provide a quick check of student understanding. Such a check might be useful to ensure that students understand prerequisite material or to provide students with a sense of accomplishment before proceeding with new material; however, the results of this experiment seemed to indicate that such questions might not lead to improved engagement or learning. Therefore, this might warrant further exploration as a factor in future experiments on clicker use. Another factor that might warrant further investigation is the type of feedback provided to students after each clicker question. For example, an instructor could simply show the bar graph of student responses, which allows students to gauge their level of understanding without taking much class time. Alternatively, an instructor could allow time for discussion—either instructor led or among small peer groups—as to why each answer was correct or incorrect. Again, this goes back to the pedagogical intent behind the use of clickers, which the results of this experiment indicate is much more important than the technology itself.

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.

References

Auras, R., and Bix, L. (2007), “WAKE UP! The Effectiveness of A Student Response System in a Large Packaging Class,” *Packaging Technology and Science*, 20, 183-195.

Beatty, I. (2004), “Transforming Student Learning With Classroom Communication Systems,” *Educause Center for Applied Research, Research Bulletin*, 2004, 2-13.

Beatty, I., Gerace, W., Leonard, W., and Dufresne, R. (2006), “Designing Effective Questions for Class-Room Response System Technology,” *American Journal of Physics*, 74, 31-39.

Beekes, W. (2006), “The ‘Millionaire’ Method for Encouraging Participation,” *Active Learning in Higher Education*, 7, 25-36.

Brewer, C. (2004), “Near Real-Time Assessment of Student Learning and Understanding in Biology Courses,” *Bioscience*, 54, 1034-1039.

Bunce, D., VandenPlas, J., and Havanki, K. (2006), “Comparing the Effectiveness on Student Achievement of a Student Response System Versus Online WEBCT Quizzes,” *Journal of Chemical Education*, 83, 488-493.

Caldwell, J. (2007), “Clickers in the Large Classroom: Current Research and Best-Practice Tips,” *CBE Life Sciences Education*, 6, 9-20.

Carnaghan, C., and Webb, A. (2006), “Investigating the Effects of Group Response Systems on Student Satisfaction, Learning and Engagement in Accounting Education,” Available at http://ssrn.com/abstract=959370.

Chen, Y., Liu, C., Yu, M., Chang, S., Lu, Y., and Chan, T. (2005), “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*, pp. 96-103.

Cleary, A. (2008), “Using Wireless Response Systems to Replicate Behavioral Research Findings in the Classroom,” *Teaching of Psychology*, 35, 42-44.

Conoley, J., Moore, G., Croom, B., and Flowers, J. (2006), “A Toy or a Teaching Tool? The Use of Audience-Response Systems in the Classroom,” *Techniques*, October 2006, 46-48.

Crossgrove, K., and Curran, K. (2008), “Using Clickers in Nonmajors- and Majors-Level Biology Courses: Student Opinion, Learning, and Long-Term Retention of Course Material,” *CBE-Life Sciences Education*, 7, 146-154.

delMas, R., Garfield, J., Chance, B., and Ooms, A. (2006), “Assessing Students' Conceptual Understanding After a First Course in Statistics,” Paper presented at the *Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Demetry, C. (2005), “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*.

Dill, E. (2008), "Do Clickers Improve Library Instruction? Lock in Your Answers Now," *Journal of Academic Librarianship*, 34, 527-529.

Duncan, D. (2005), *Clickers in the Classroom: How to Enhance Science Teaching Using Classroom Response Systems*, Pearson, San Francisco, CA.

Fagan, A., Crouch, C., and Mazur, E. (2002), "Peer Instruction: Results from a Range of Classrooms," *The Physics Teacher*, 40, 206-209.

Fredricks, J., Blumenfeld, P., and Paris, A. (2004), "School Engagement: Potential of The Concept, State of the Evidence," *Review of Educational Research*, 74, 59-109.

Freeman, S., O'Connor, E., Parks, J., Cunningham, M., Hurley, D., Haak, D., Dirks, C., and Wenderoth, M. (2007), "Prescribed Active Learning Increases Performance in Introductory Biology," *CBE Life Sciences Education*, 6, 132-139.

Greer, L., and Heaney, P. (2004), "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.

Hanley, J., and Jackson, P. (2006), "Making It Click: A California *High* School Test Drives and Evaluates Six New Personal Response Systems," *Technology and Learning*, 26, 34-38.

Jackson, M., and Trees, A. (2003), "Clicker Implementation and Assessment," Available at http://comm.colorado.edu/~jackson/clickerreport.htm.

James, M. (2006), "The Effect of Grading Incentive on Student Discourse in Peer Instruction," *American Journal of Physics*, 74, 689-691.

Kennedy, G., and Cutts, Q. (2005), "The Association Between Students' Use of an Electronic Voting System and Their Learning Outcomes," *Journal of Computer Assisted Learning*, 21, 260-268.

Lass, D., Morzuch, B., and Rogers, R. (2007), "Teaching With Technology to Engage Students and Enhance Learning," Working Paper No. 2007-1. Available at http://ssrn.com/abstract=958036.

Latessa, R., and Mouw, D. (2005), "Use of an Audience Response System to Augment Interactive Learning," *Family Medicine*, 37, 12-14.

MacGeorge, E., Homan, S., Dunning Jr., J., Elmore, D., Bodie, G., Evans, E., Khichadia, S., Lichti, S., Feng, B., and Geddes, B. (2008), "Student Evaluation of Audience Response Technology in Large Lecture Classes," *Educational Technology Research and Development*, 56, 125-145.

Mayer, R., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., and Zhang, H. (2009), "Clickers in the College Classroom:

Fostering Learning With Questioning Methods in Large Lecture Classes," *Contemporary Educational Psychology*, 34, 51-57.

Mazur, E. (1997), *Peer Instruction: A User's Manual*. Prentice Hall, Upper Saddle River, NJ.

Miller, R., Ashar, B., and Getz, K. (2003), "Evaluation of an Audience Response System for the Continuing Education of Health Professionals," *Journal of Continuing Education of Health Professionals*, 23, 109-115.

Morling, B., McAuliffe, M., and Cohen, L. (2008), "Efficacy of Personal Response Systems ("Clickers") in Large, Introductory Psychology Classes," *Teaching of Psychology*, 35, 45-50.

Nelson, M., and Hauck, R. (2009), "Clicking to Learn: A Case Study of Embedding Radio-Frequency Based Clickers in an Introductory Management Information Systems Course," *Journal of Information Systems Education*, 19, 55-64.

Nosek, T., Wang, W., Medvedev, I., While, M., and O'Brian, T. (2006), "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*, pp. 2245-2250.

Nunnally, J. (1978), *Psychometric Theory*, McGraw-Hill, New York.

Penuel, W., Boscardin, C., Masyn, K., and Crawford, V. (2007), "Teaching With Student Response Systems in Elementary and Secondary Education Settings: A Survey Study," *Educational Technology Research and Development*, 55, 315-346.

Pradhan, A., Sparano, D., and Ananth, C. (2005), "The Influence of an Audience Response System on Knowledge Retention: An Application to Resident Education," *American Journal of Obstetrics and Gynecology*, 193, 1827-1830.

Preszler, R., Dawe, A., Shuster, C., and Shuster, M. (2007), "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.

Rogers, R. (2003), "Using Personal Response Systems to Engage Students and Enhance Learning," Presented at *Making Statistics More Effective in Schools and Business (MSMESB) Conference*.

Roselli, R., and Brophy, S. (2006), "Experiences With Formative Assessment in Engineering Classrooms," *Journal of Engineering Education*, 95, 325-333.

Schackow, T., Chavez, M., Loya, L., and Friedman, M. (2004), "Audience Response System: Effect on Learning in Family Medicine Residents," *Family Medicine*, 36, 496-504.

Schau, C. (2003), "Students Attitudes: The "Other" Important Outcome in Statistics Education," Paper presented at the *American Statistical Association Joints Statistical Meetings*, Alexandria, VA.

Schau, C., Stevens, J., Dauphinee, T., and Del Vecchio, A. (1995), “The Development and Validation of the Survey of Attitudes Toward Statistics,” *Educational and Psychological Measurement*, 55, 868-875.

Siau, K., Sheng, H., and Nah, F. (2006), “Use of a Classroom Response System to Enhance Classroom Interactivity,” *IEEE Transactions on Education*, 49, 398-403.

Stowell, J., and Nelson, J. (2007), “Benefits of Electronic Audience Response Systems on Student Participation, Learning and Emotion,” *Teaching of Psychology*, 34, 253-258.

Trapskin, P., Smith, K., Armitstead, J., and Davis, G. (2005), “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.

Uhari, M., Renko, M., and Hannu, S. (2003), “Experiences of Using an Interactive Audience Response System in Lectures,” *BMC Medical Education*, 3.

Van Dijk, L., Van Den Berg, G., and Van Keulen, H. (2001), “Interactive Lectures in Engineering Education,” *European Journal Of Engineering Education*, 26, 15-28.

Wit, E. (2003), “Who Wants To Be... The Use of a Personal Response System in Statistics Teaching,” *MSOR Connections*, 3, 14-20.

Zhu, E. (2007), “Teaching With Clickers,” CRLT Occasional Paper No. 22. Available at http://www.crlt.umich.edu/publinks/CRLT_no22.pdf.

Zualkernan, I. (2007), “Using Soloman-Felder Learning Style Index to Evaluate Pedagogical Resources for Introductory Programming Classes,” In *Proceedings of the Twenty-ninth Inter-national Conference on Software Engineering*, pp. 723-726.