



# Planning a Comparative Experiment in Educational Settings

[Herle M. McGowan](#)

North Carolina State University

*Journal of Statistics Education* Volume 19, Number 2 (2011),  
[www.amstat.org/publications/jse/v19n2/mcgowan.pdf](http://www.amstat.org/publications/jse/v19n2/mcgowan.pdf)

Copyright © 2011 by Herle M. McGowan all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

---

**Key Words:** Quantitative research; Efficacy; Evaluation; Statistics education; Clickers.

## Abstract

A well-designed experiment is the best method for establishing efficacy of any intervention, be it medical, behavioral, or educational in nature. This paper reviews the steps necessary in conducting a comparative experiment in an educational setting, and illustrates how these steps might be fulfilled within the context of a large-scale randomized experiment conducted in an introductory statistics course. The primary goal of this paper is to help researchers identify salient issues to consider and potential pitfalls to avoid when designing a comparative experiment in an educational setting.

## 1. Introduction

Educational researchers in many disciplines are faced with the task of exploring how students learn and are correspondingly addressing the issue of how to best help students do so. Often, educational researchers are interested in determining the effectiveness of some technology or pedagogical technique for use in the classroom. Their ability to do so depends on the quality of the research methodologies used to investigate these “treatments.”

It is commonly known that a well-designed randomized experiment is the best method for establishing efficacy of any intervention, be it medical, behavioral, or educational in nature. While the use of randomized, comparative experiments in educational settings is not without criticism (e.g. [Howe 2004](#); [Cook 2002](#)), it does play a role in each phase of an educational research program, from studies of initial efficacy to larger trials that confirm or optimize effects of educational interventions ([SMER 2007](#)). However, a review of comparative studies

conducted in the field of statistics education indicates that the methodology currently used could be improved ([McGowan 2009](#)).

This paper reviews the steps necessary in conducting a comparative experiment. The paper illustrates how these steps might be fulfilled within the context of a large-scale randomized experiment conducted in an introductory statistics course. The primary goal of this paper is to help researchers identify salient issues to consider and potential pitfalls to avoid when designing a comparative experiment in an educational setting.

## 2. The Experimental Process

Most textbooks that discuss designing an experiment—for classroom research or in other contexts—discuss the same basic steps (see, for example, [Slavin 1984](#); [Light, Singer and Willett 1990](#); [Wu & Hamada 2000](#)). The first step involves specification of the problem, including defining the research question or hypothesis to be tested, and also identifying the response and treatment variables or other predictors of interest. After these are specified, extensive planning of the experimental design and procedures is necessary, including selection of measures, participants, and a plan for randomization. Planning is followed by implementation, analysis, and finally, drawing conclusions from the data. In this section, several of the choices available to a researcher for each of these steps are reviewed. In Section 3, these choices and their associated consequences are illustrated within the context of an example—an experiment exploring the use of personal response systems in teaching statistics.

### 2.1 Specification of the research problem

#### 2.1.1 Defining the research question or hypotheses

The first step in any research study is to define the question or hypothesis of interest. Motivation could come from an interest in a particular treatment, such as wanting to explore the effects of a new technology, or in a particular outcome, such as seeking ways to improve understanding of a concept with which students commonly struggle. Whatever the motivation, each research problem should be well-grounded in current understanding of how students learn.

The scope of each research question should be appropriate given the current state of knowledge in a particular field. For example, if a treatment has not been extensively studied, questions exploring basic efficacy are the necessary starting point. However, as knowledge begins to grow, the nuance of research questions should also grow. For example, questions of efficacy could be followed by questions that explore particular aspects or variations of the treatment, to learn *how* or *why* it is successful, or under what conditions it is optimal.

Finally, in order to have a viable research problem, ideas need to be focused into a narrow, specific question that can be clearly answered. [Garfield \(2006, p. 8\)](#) provides this example to demonstrate the difference between a broad and a specific research question: “Does technology improve student learning?” could be focused into “How can a particular use of a specific technology tool help students understand the meaning of confidence interval?”

### **2.1.2 Identifying the outcome and treatment variables**

The process of specifying the research question and hypotheses will help delineate what are the relevant outcome and treatment variables (in fact, these steps are often intertwined). As with specification of the research question, defining the treatment variables and outcomes should be supported by a thorough review of the literature to determine current understanding of best practice. This can help with selection of the best treatments to explore, as well as with determining the most appropriate way to measure the outcomes. Another approach is to consider the four levels of program evaluation described in [Kirkpatrick and Kirkpatrick \(2006\)](#). Consideration of the first three levels can help with specification of the outcome variable(s) by identifying 1) the desired results of treatment 2) the behaviors necessary to achieve these results, and 3) the attitudes, knowledge or skills that could produce the desired behaviors. The fourth level deals with details of implementation—how to present the intervention to participants so that they react favorably to it. Details of implementation are discussed in Section 2.3.

## **2.2 Planning the experimental design and procedures**

A major consideration in designing any experiment in an educational setting is to ensure that experimental procedures are 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 “guinea pigs.” As researchers, we are under the governance of Institutional Review Boards (IRBs), which make sure that students’ rights are protected. Additionally, planning experimental procedures as a normal part of a course, to the extent possible, makes implementing the experiment easier since special treatment is not required for students who do not wish to participate. Some considerations in planning a non-obtrusive experiment are discussed throughout this section.

### **2.2.1 Measuring the outcomes**

The selection of high-quality assessment instruments is important for getting good data in any study. Instruments should produce data that are valid (i.e. measuring what it intends to measure) and reliable (i.e. measuring consistently) ([Nunnally 1978](#)). There is an entire academic field dedicated to the science of developing valid and reliable instruments, a process which takes time and refinement. Often assessments used in the normal process of a course, such as an instructor developed exam or survey, will not achieve these properties. Use of standardized assessments is preferable, as these have been through testing and refinement. Additionally, use of a nationally available instrument helps frame the results of the present research (e.g. Do higher scores indicate better conceptual understanding or better procedural ability?), allows for easier comparison of results across studies using the same measure, and allows for easier reproduction of experimental conditions in future studies. This, in turn, facilitates building a body of knowledge about a particular treatment or a particular outcome. To ensure that students are not over-burdened by assessment, standardized instruments could be incorporated into typical course assessment; for example, in place of all or part of an instructor developed exam.

### 2.2.2 Measuring the treatment variables

Development of an operational definition of the treatment variables—something that can actually be carried out in a study—will be heavily influenced by the particulars of the treatment being investigated. For example:

- What technology will be available to the instructor and/or students with which to implement the treatment (if necessary)?
- How much time is available for the treatment? This could range from a few minutes for a single activity to an entire semester.
- How much treatment (dosage) is appropriate, or is possible to implement given constraints (e.g. time, resources, workload)?
- How many levels of treatment are needed? For example, if the research question is of the form “Is treatment better than no treatment?” then two levels (e.g. ‘some’ vs. ‘none’) are needed. If the question is of the form “How much treatment is best?” then more than two levels (e.g. ‘high’ vs. ‘moderate’ vs. ‘low’) may be warranted.

The number of levels at which a treatment is measured will have direct impact on the design of the experiment. The simplest treatment in a comparative study has two levels (e.g. ‘some’ vs. ‘none’; ‘more’ vs. ‘less’) and is investigated with a 2-group design. More complex treatments warrant more complex designs. For example, factorial designs, which are common in industrial experiments and are especially well suited to investigate interactions, could be used to explore several treatment variables ([Wu & Hamada 2000](#)) at two or more levels each. Factorial designs could be useful in educational research for exploring the optimization of a particular treatment after initial efficacy of that treatment has been established.

### 2.2.3 Selection of participants

The identification of appropriate participants and comparison groups will be dictated by the research problem and the treatment of interest. Often, the pool of eligible participants will be all students registered for a particular class. Research conducted at a university that involves human subjects will likely necessitate approval from an IRB, and the researcher will have to seek student consent to participate in the study. Depending on the nature of the treatment, those students who do not wish to participate may need to be separated from those who do. However, this could be avoided if experimental procedures are designed to be an integral part of course activities, so that all students participate in the activities during class or as part of required out-of-class work. In this case, consent should be sought to analyze and publish results based on the students’ data (see [Appendix A](#) for an example consent form).

### 2.2.4 The use of randomization

The use of randomization will be dictated by practical and ethical concerns. Of course, randomization of individual students is the best procedure to ensure baseline equivalence of the comparison groups; however, this is not always possible. It may be easier to randomize individual students when treatment is something that takes place on an individual level or is something that represents a small part of the course, such as individual exploration of a particular concept using a computer applet. In contrast, treatment may be delivered to entire class sections

over the full course of a semester. In cases such as this, it may be difficult to randomly assign individual students due to scheduling constraints, especially if sections are offered on multiple days and times, as is common in large college courses. Seeking volunteers to be randomized among particular sections or time slots would greatly reduce the number of participants. Another alternative is to randomize entire classes to treatment conditions. In this case, it is best to randomize multiple sections to each comparison group so that the effect of a treatment variable is not confounded with group factors. This will have implications for how the resulting data is analyzed, as discussed in Section 2.4.

In addition to potentially easier randomization, there are other benefits to having a small, focused treatment as opposed to one that encompasses an entire course or semester. For example, this type of treatment is more consistent with the recommendation of asking narrow, focused research questions made in Section 2.1. This type of treatment also requires less time, money, and effort to implement, whereas implementation of a more complex treatment can be difficult. Of course, there is a risk that a small treatment could be associated with a small effect, which would in turn be difficult to detect. Strong connection to current theory at the point of specifying the research problem can help maximize the potential effect of a treatment; likewise careful selection and placement of assessments can maximize ability to detect the effect.

### **2.3 Implementation of the experiment**

The process of planning the implementation of an experiment will likely go through several stages. There are many things to consider, including the ideal implementation that would be necessary to exactly answer the research question of interest (e.g. randomizing individual students to better make causal claims about the success of the treatment), as well as the implementation that is actually possible given constraints due to time, money, administrative oversight and the like. Many IRBs will require a detailed plan for implementation before they approve a project.

In addition to having a plan for implementation, it is important to have a clear way to communicate this plan to additional instructors who may be implementing the intervention. For example, weekly meetings or memos with directives could be used to ensure consistency in experimental procedures between multiple classrooms. Still, deviations from the plan are bound to occur during any experiment. Thus, it is a good idea to maintain records of actual implementation in addition to planned implementation. Having such records not only reveals infidelity, but also provides some idea of how often such problems occur. This information can be useful for evaluating the results of an experiment and explaining why they may or may not be as expected. Routinely reviewing implementation records during the experimental period provides the principle investigator a chance to correct problems during the current experiment, or to refine the treatment or the implementation plans for future replications.

### **2.4 Analysis and conclusions of the experiment**

With any study, features of the design will have an impact on what analysis is appropriate. In educational research, two common design features that will affect analysis are lack of randomization of individual students to treatment conditions and the delivery of treatment to an

entire classroom of students. The fact that students are in the same class, or that multiple classes may be taught by the same instructor, violates the condition of independence required by standard statistical models ([SMER 2007](#)). Hierarchical, or multi-level, models with nested random effects should be used to address this issue, and recent advances in software have made this type of analysis much more user friendly (see, for example, [Pinheiro and Bates 2009](#); [Raudenbush and Byrk 2002](#)).

The use of group randomization can also cause covariate imbalances prior to the start of treatment that will need to be accounted for in the modeling process. In the planning stage of the experiment, potentially important covariates should be identified and subsequently measured. Identification of such covariates could occur through a review of previous research or through discussions with fellow researchers or instructors.

The final step of the experimental process is to consider the limitations of a study's conclusions, which again often relate back to the choices made in the design and planning of the experiment. Honest presentation and discussion of limitations can be useful for planning replicating or similar experiments, which in turn helps build the body of knowledge in a field.

When it comes to publishing about the experiment, report as much detail about the design, implementation (the plan and any deviations from that plan), and results (including descriptive statistics, test-statistics and  $p$ -values, as well as confidence intervals or effect sizes) as is possible given space constraints. The [SMER report \(2007\)](#) provides an extensive list of what should be reported, with the goal of allowing future replication of the experiment.

### **3. An Illustrative Example**

The experimental process described in the previous section will now be illustrated through an example—an experiment that was conducted at a large, mid-western research university. This particular experiment was fairly complex, which demonstrates that implementation of rich experiments is possible in educational settings, and sometimes even necessary to advance knowledge. For simplicity, the information presented here represents only a portion of the full experiment; however what is presented is still fairly detailed, to indicate the level of thought that can go into the planning process. While not every comparative study needs to consider each issue at the level presented here, it is hoped that presentation of such detail will help new educational researchers consider issues they might otherwise not have (indeed, not every issue was evident to the author during the planning of this experiment; some only became so in hindsight). Finally, discussion of what could have been done differently is presented throughout this section, to illustrate the reflection on the design and implementation of an experiment that should take place after its completion.

The general purpose of the illustrative experiment was to explore the effectiveness of personal response systems, or “clickers,” as a pedagogical tool in statistics. Clickers are hand-held remotes that allow students to respond to questions, usually multiple-choice, posed by the instructor during a class. Software then collects and tallies these responses almost instantly. A bar-graph of the frequency of each answer choice can be displayed to students, allowing them to see if they were correct or not. Many papers discuss the use and potential benefits of clickers in

the classroom; readers interested in learning more about this technology are referred to summaries of this literature (e.g. [Duncan 2005](#); [Caldwell 2007](#); [Zhu 2007](#)) and to guides on writing good clicker questions (e.g. [Beatty 2004](#); [Beatty, Gerace, Leonard and Dufresne 2006](#)).

The clicker experiment was implemented in a multi-section introductory statistics course for college undergraduates. The course included 80 minutes of lab practice in addition to three hours of lecture per week. The purpose of lecture was to introduce the bulk of the course material, with students then being able to apply their knowledge during lab. Lecture sections varied greatly in terms of their size, the number of sessions per week, and the length of each session. Labs, on the other hand, were fairly uniform with respect to these aspects: there were about 25 students in each lab section, which met once a week for 80 minutes. There were also many more lab sections than lecture sections (fifty compared to six). For these reasons the experiment was implemented in the lab sections of the course. More details on the design and implementation of the experiment will be provided throughout this section; however detailed analysis and results of the study have been published elsewhere ([McGowan and Gunderson 2010](#)).

### **3.1 Specification of the research problem**

#### **3.1.1 Defining the research question or hypotheses**

The research problem for the clicker experiment arose out of the natural process of improving the course. The lead instructor felt that there would be benefits to the clicker technology and thus began using them. In informal terms, the research problem for the experiment was to investigate if some uses of clickers were better than others. The process of formalizing this research question is specified in the next subsection.

#### **3.1.2 Identifying the outcome and treatment variables**

The outcomes for the clicker experiment were “engagement” and “learning”; identifying these followed naturally from the decision to study clickers, as engagement and learning are widely believed to be the benefits of any educational technology. A review of relevant literature was then used to help define and operationalize these outcomes (see Section 3.2.1).

“Some uses of clickers” was formalized by selecting three particular aspects of clicker use that were believed to affect engagement and learning. In the literature on clickers, 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 author’s experience in teaching with clickers seemed to indicate that there might be practical limits as to how to provide this feedback. Clickers were first introduced in the course during days of exam review in labs. Students were given an opportunity to work on review problems in groups and then click in the answers to several problems in succession. During these sessions, students often became distracted and began talking or looking online while waiting for others to enter their answer to a question. This could be indicative of a negative interaction between the number of clicker questions asked and how those questions were incorporated into the class session. The possibility of an ‘overdose,’ so to speak, of clicker use had not been widely considered, so the

experiment was designed in part to address this gap in the literature. To that extent, two of the treatment variables considered in the experiment were the number of questions asked with clickers during a lab session (called *Frequency*) and the placement of those questions throughout the material (specifically, if the questions were asked in a group or more spread out; called *Agglomeration*). Measurement of each of these treatment variables is described in Section 3.2.2.

The specific research questions for the clicker experiment were then formalized as:

1. What is the main effect of *Frequency*?
2. What is the main effect of *Agglomeration*?
3. Is there a negative interaction between *Frequency* and *Agglomeration*?

The scope of these research questions was appropriate given the knowledge of clickers at the time. Several studies had explored the efficacy of this technology and found evidence that it was beneficial for students. The research questions in the current experiment were selected to add to the knowledge about clickers by exploring factors that had not been extensively studied and that might contribute to optimal use of clickers in the classroom.

## 3.2 Planning the experimental design and procedures

### 3.2.1 Measuring the outcomes

The clicker experiment relied extensively on standardized assessments to measure engagement and learning. For example, the Survey of Attitudes Towards Statistics (SATS; [Schau, Stevens, Dauphinee and Del Vecchio 1995](#)) was used in part to measure engagement, and several instruments from the Assessment Resource Tools for Improving Statistical Thinking project (ARTIST; <https://app.gen.umn.edu/artist/>) were used to measure learning, including the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS; [delMas, Garfield, Chance and Ooms 2006](#)) and four topic-specific scales (Normal Distribution, Sampling Distributions, Confidence Intervals, and Hypothesis Testing). CAOS served as a comprehensive assessment of statistical understanding both at the beginning and end of the experiment. The topic scales served as more proximal measures of understanding about particular topics. The topic scales were administered at equal increments over the semester, after the presentation of the corresponding material in lecture. This avoided excessive testing that may have occurred if many assessments were administered in a short amount of time.

To ensure that the assessment process was not too burdensome for students, assessments were administered during class time; this also ensured higher completion rates. Typically, assessments were completed at the beginning of a class in hopes of decreasing the urge to rush through just to get it over with and get out the door. Since assessments were a part of class time, students were awarded participation points for completing them. Additionally, the instruments were selected to provide more than a score for the purposes of the experiment alone. It was hoped that these instruments would help increase students' broad conceptual understanding and also provide formative feedback as to their level of understanding—before losing points on homework assignments or exams.

For ease of implementation, each of the outcome measures were administered online, using software maintained by the university. 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. It also allowed for the order of the questions and their answer choices to be randomized for individual students. The 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 or, if the instructor desires, to view the questions or correct answers after submission. Data were securely backed-up on university servers. Data could be outputted in several formats for exploration and analysis. While the particular software used was specific to the university, similar services may be available at other universities. Additionally, commercial learning management systems, such as Moodle or Blackboard, could be used. Certainly the use of online data collection is not new, but it is worth noting that implementation of an experiment of this size and complexity would not have been feasible without it.

While there were several good aspects to the measurement of outcomes in the clicker experiment, a limitation of their use was noticed. Specifically, feedback from students revealed that they did not consider the questions on CAOS or the topic scales to be in line with questions on homework and exams, but instead saw these assessments as disjoint from the rest of the coursework. While the instruments were chosen specifically for their focus on conceptual issues—something that students often struggle with—many homework and exam questions were problem-solving or procedurally based. The perception that these assessments did not “fit” with the rest of the course, coupled with the fact that their impact on a student’s course grade was through completion rather than correctness, may have lead to students not trying very hard on these assessments. This in turn could mean that the resulting scores are not a good reflection of student understanding. In future experiments, this could be avoided by better incorporating the assessments into the course, for example as part of a course exam.

### 3.2.2 Measuring the treatment variables

The first treatment variable, called *Frequency*, considered the number of clicker questions asked during a class. This variable was measured at two levels: *High* (at least six clicker questions were asked) and *Low* (3-4 clicker questions were asked). The second treatment variable, called *Agglomeration*, considered the placement of the questions throughout the material. This variable was also measured at two levels: *On* (clicker questions were asked in an agglomerate or group) and *Off* (clicker questions were dispersed throughout the session). Selection of these levels was influenced by practical issues, such as making sure the resulting combinations of levels would make sense. For example, it was decided that asking two clicker questions in a row was not excessive, and might actually be very useful for reinforcing concepts by asking a follow-up question. Therefore, three questions were considered the minimum number to define an “agglomeration” of questions. Three questions was also set as the lower bound for the *Low* level of *Frequency* because, otherwise, the combination of asking fewer than three clicker questions in an agglomeration would not have been possible. The lower bound for the *High* level could have been set at five clicker questions, but having a distinct gap—albeit a small one—between the levels makes it easier to detect any difference that might exist between them.

The clicker questions themselves were taken directly from existing questions in the students' lab workbook, so that no extra material was added into already full lab periods. Using questions that would have been asked anyway ensured that clicker use was seamlessly integrated into labs, increasing the intrinsic value of the questions and the clickers themselves (meaning that clicker use was a component of the course, not something added in solely for the purpose of the experiment that students do not have to take seriously). Finally, this made it easier for the same questions, with the same answer choices, when appropriate, to be asked in every lab section. The sections differed with respect to the number of questions asked using clickers and the placement of the clicker questions within the lesson (whether those questions were grouped together or not). This avoided 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 ([Carnaghan and Webb 2006](#)).

It is worth noting here that, in a simpler version of this experiment, either of the treatment variables could have been investigated in isolation. This would have resulted in a two-group comparison that would be possible to implement in a smaller course. Similarly, only one of the outcomes could have been measured, which would have reduced the time and resources needed for data collection.

### **3.2.3 Selection of participants**

All students who were at least 18 years of age and were registered in the course after the university's add/drop deadline were eligible to participate in this experiment. Waiting until after this deadline avoided having to deal with turnover in student enrollment early in the semester (which may be common in large service courses). Since experimental procedures were designed to be an integral part of course activities—meaning that all students completed the activities as part of their course grade—we did not need to seek student consent to be a part of these activities or separate those who wished to participate from those who did not. Instead, students provided permission for their data to be analyzed (see [Appendix A](#)).

### **3.2.4 The use of randomization**

In the clicker experiment, the unit of randomization was the lab instructor, not the lab section or the individual students themselves. Students, who had no prior knowledge of the experiment, were allowed to register for any section of the course. Each lab instructor—who taught 2-3 sections—was then randomized to a treatment condition, so that all of their sections and students would be under the same condition. This was done to make things simpler for the lab instructor, also hopefully limiting ‘contamination’ between treatment groups that could result from a lab instructor confusing sections. However, this did have implications for how the resulting data were analyzed, as discussed in Section 3.4.

### 3.3 Implementation of the experiment

Planned implementation procedures were communicated to instructors through weekly meetings and memos, which were already used in the course to help ensure consistency in teaching and grading among the fifty lab sections. During these meetings, the lab instructors and the lead instructor discussed what did or did not go well in the previous lesson, addressed questions about grading the homework, and went over the lesson plan for the coming week. During the experimental semester, the principle investigator also discussed implementation of the experimental conditions for the coming week. The weekly memo included the meeting agenda as well as a schedule of specific activities to cover in the following lab. During the experimental semester, the memos for the weekly meetings were personalized for each lab instructor. Memos were color coded based on the lab instructor's assigned treatment group (e.g. the treatment condition with *Frequency* at the *Low* level and *Agglomeration* set to *Off* was referred to as the "Blue Team," and all lab instructors assigned to this group knew to look for their memo on blue paper). Additional information, such as the lab instructor's name and other personalized instructions, were included at the top of the page using a simple mail merge feature in a word processing software. Appendix B shows an example of one of these weekly memos for an anonymized instructor.

Actual implementation in the clicker experiment was tracked using a half-page survey, which lab instructors were asked to fill out after each lab (see [Appendix C](#)). This survey asked them to report the levels of each treatment variable that had been received by the class and the number of students in attendance (used to assess the proportion of students using clickers). The survey also asked general questions about the existence of technical or other difficulties during lab and reminded lab instructors to upload the clicker response files to a central database for the principle investigator. This survey was used to identify and correct problems with implementation.

This survey was also used to evaluate the subsequent results of the experiment. For example, there were inconsistencies in the specific placement of individual clicker questions within a class period. Lab instructors had been provided with some guidance as to how to incorporate clicker questions into lab (e.g. to ask all questions at the end of an activity or to incorporate the questions into the activity). However, specific instructions, which might restrict the lab instructors' teaching, were kept to a minimum to avoid conflicts in the team or with the experimental procedure. In hindsight, the general guidance provided as to the placement of clicker questions was not enough. Lab instructors varied in their interpretation of this guidance and their ultimate placement of the questions. It was not always clear to lab instructors, 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 may simply require recall ability when asked after. It is believed that this in turn affected the ability to detect any treatment effects of *Frequency* and *Agglomeration*. 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 (one for each possible combination of the levels of *Frequency* and *Agglomeration*) over nine weeks, and in terms of excessive reduction of the lab instructors' freedom in teaching. In

conversations with lab instructors 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, making the treatment smaller and more focused, which in turn might make more extensive scripting and lab instructor training feasible.

### **3.4 Analysis and conclusions of the experiment**

In the clicker experiment, group randomization was used to assign instructors to treatment conditions. As such, hierarchical, or multi-level, models were used for each analysis conducted. These models included random effects for students nested within lab, which were in turn nested within lab instructor. Also, to account for covariate imbalances between treatment groups, each model adjusted for important confounding variables.

Again, it should be noted that the specific results from the clicker experiment are published elsewhere ([McGowan and Gunderson 2010](#)). Considering the results—what factors were and were not significant—and the implementation of the experiment lead to some important findings about what could have been improved if this were to be repeated. For example, the decision to implement the treatment in labs rather than lectures had unintended consequences on the results of the experiment. As has been mentioned before, lab sections were more plentiful in number and more uniform in terms of size than the lecture sections. The consistent schedule of lab once a week for 80 minutes—with the exact same activities covered in each section—was much more conducive to the implementation of the experimental design. However, the very purpose of the labs was 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.

Considering the limitations of this experiment also led to ideas for future research on clickers. For example, an aspect of clicker use that was not studied explicitly in this experiment, but in hindsight appeared to be extremely important, was that of question purpose. Many questions in this experiment involved factual recall, which could be useful for ensuring that everyone in the class understands required material. Fewer questions involved applying or extending concepts in the low-stakes, instant feedback environment afforded by the clicker technology. Future experiments could explore this distinction to determine which purpose is more beneficial for students, or under which circumstances each is most appropriately used. Related to this could be the factor of what instructors do with the instant feedback provided by the clickers. Do they simply tell the correct answer and move on? Lecture on why each response is or is not correct? Allow for class discussion or activities to explore the concept further? Clearly, there is still much to be learned about clickers as an educational technology; honest reflection on each study about clickers can help connect and ultimately expand this knowledge.

## **4. Summary**

This paper reviewed the necessary steps in conducting a comparative experiment and discussed some of the decisions that need to be made by an educational researcher at each step. The guidance provided throughout the paper included:

- Begin every experiment with a literature search to explore what is known about the research problem, treatment variables, and outcomes of interest. Use this literature to guide the decisions made in planning the design and implementation of the experiment.
- Questions of initial treatment efficacy should be followed-up with questions that allow for identification of the “active” ingredient(s) in the success of a treatment, so that ingredient could possibly be replicated in future experiments. Multifactor designs, such as factorial designs, could be used to explore and refine a complex treatment.
- Pretreatment differences, which could arise due to group assignment or group delivery of treatment, need to be accounted for. This can be done through design (e.g. by randomizing multiple sections to each treatment condition) and analysis (e.g. through covariate adjustment).
- Use valid and reliable assessment instruments when measuring outcomes, particularly learning outcomes. Standardized assessments of learning in statistics, such as the CAOS test, already exist and could easily be incorporated as part or all of a course exam.
- Use hierarchical modeling to analyze nested data. Given that nearly every educational intervention is implemented on groups of students nested within a classroom that is nested within a school, nearly every analysis in education should be hierarchical.
- Have a detailed plan for implementation, and keep records of deviations from this plan. Be as detailed as possible (given space constraints) when describing the design and implementation of an experiment, as this will facilitate building a body of knowledge about a treatment or an outcome.

Finally, a few points of pragmatic advice:

- Not all experiments need to be as complex as the clicker experiment presented here. Starting with something small is better than doing nothing at all, and could provide a foundation for future research.
- Seek help or advice when planning any experiment, whether it is large or small. If there are not colleagues within your own department that could help, you could look in other departments or at other institutions. Additionally, the research arm of the Consortium for the Advancement of Undergraduate Statistics Education ([causeweb.org](http://causeweb.org)) offers resources which may be of use when planning a research study.
- Make use of resources that are available to you. For example, learning management software that may already be used for a course could also be used for data collection. Your institution may have funds available for research on teaching and learning or for course development that could be used to start a project, possibly even funding a student to help with logistics or data management.
- Finally, automate whatever you can, such as data collection, assessment scoring, or communication with any other implementers.

Planning an experiment in any setting requires a great deal of thought and careful consideration—this is especially true when planning an experiment in an educational setting. The nature and structure of education provides additional complexities in the experimental process, as have been discussed throughout this paper. However, it is possible to conduct a well-designed experiment in a classroom. If done with care and a strong connection to previous research, we can make great gains in our understanding of how students learn and how to best facilitate that process.

## **Appendix A**

### **Informed Consent Document**

#### A Study on the Effectiveness of Clickers in the Statistics Classroom

You are invited to be part of a research study on the effectiveness of clickers in helping to engage students in the Statistics classroom and learn the subject. You were selected as a possible participant because you are enrolled in [Insert course name]. We ask that you read this form and ask any questions you may have before deciding to participate in the study.

This study is being conducted by: [Insert primary investigator name and affiliation]

#### Background:

The purpose of this study is to assess the effectiveness of clickers in helping to engage students in statistics classrooms and learn the subject. Some people believe that using clickers helps to engage the students and hence improves the learning experience in the class. One of our main goals is to test this hypothesis. If using the clickers leads to an improvement, we want to learn about the best ways to use clickers, including how frequently they should be used and when.

#### Procedures:

Agreeing to participate does not require you to complete any work beyond normal course requirements. Participation in this study means that you provide permission to use the data we collect from surveys, clicker responses, in-lab reviews and other assessments in the research project. Your responses will be combined with those of other participants and reported in aggregate form. Information about individual students will not be used in any published reports.

#### Risks and Benefits of being in the Study:

There is no risk in participating in this project. Although you may not receive direct benefit from your participation, others may ultimately benefit from the knowledge obtained in this study.

#### Compensation:

You will receive compensation for the work you complete, in the form of class participation points. Even if you choose not to participate in this study – meaning that you do not want your data to be used in this research project – you will receive the same compensation. Refusal to participate in this project will not affect your grade.

#### Confidentiality:

The records of this study will be kept confidential to the extent provided by federal, state, and local law. However, the Institutional Review Board or university and government officials responsible for monitoring this study may inspect these records. In any reports on this study, we will not include any information that will make it possible to identify an individual student.

#### Voluntary Nature of the Study:

Your participation in this project is voluntary. Even if you sign the informed consent document, you may decide to leave the study at any time without penalty or loss of benefits to which you

may otherwise be entitled. You may skip or refuse to answer any survey question without affecting your study compensation or academic standing/record.

**Contacts and Questions:**

The researcher conducting this study is [Insert primary investigator name]. If you have questions about this study, you may contact [him/her] at [Insert contact information, e.g. address, phone number, email]. Should you have questions regarding your rights as a research participant, please contact the Institutional Review Board, [Insert IRB contact information]. A copy of this document will be kept together with the research records of this study. The information contained in this document is available on the course website for your reference.

**Statement of Consent (check the appropriate boxes here):**

**Age Verification:**

I am 18 years of age or older

I am less than 18 years of age

**Consent:** I have read and understood the above information.

I agree to participate in the study.

I do NOT want to participate in the study

---

Printed Name

---

Signature

## Appendix B Sample Memo with Implementation Instructions

(Note that underlined text was inserted via mail merge and was personalized for each implementer.)

Name: Doe, Jane                      Frequency: Low                      Agglomeration: Off

Week 4: In-Lab Review of Normal Distrib., Sampling Distrib. and CLT Ideas

Before lab: Download the presentation Feb2-4 Blue.ppt

During lab:

- ~ For students that missed last week:
  - o Have them complete the Informed Consent before leaving class
  - o Have them complete the Attitudes Survey and CAOS before midnight Friday (links on course website)
- ~ For students that joined the class after the first week of labs:
  - o Have them complete the Background Info survey before Friday midnight (link on course website)
- 1. In-Lab Review on Normal Distributions (link is on course website). Time = 12-15 minutes
- 2. Do Module 4: Sampling Distributions and the CLT. Time = about 30 minutes
  - a. Start with a brief overview of sampling distributions. In particular you want to emphasize the fact that statistics calculated from random samples are also random variables, so they have their own distributions. It is important for students to understand the fact that we are studying the distribution of statistics.
  - b. Work through the first three tasks according to your assigned experiment level.
    - i. Give students a minute to work on part (a) of a problem and then ask the relevant clicker question; then give them a minute to work on part (b) before asking the relevant clicker question; continue in this fashion for all questions.
  - c. Make sure you emphasize Step 4. This is the “take away” from the simulation.

After lab...

- ~ Fill out your Lab Log. Put this in the PI’s mailbox or bring to the next GSI meeting.
- ~ Upload your results files to your drop box on the website.

## Appendix C

### Log for Recording Actual Implementation

Lab Log for (name) \_\_\_\_\_ Team \_\_\_\_\_ Week \_\_\_\_\_

1. Section number \_\_\_\_\_
2. Levels for Frequency/Agglomeration:    Low/Off    Low/On    High/Off    High/On
3. How many clicker questions did you ask? \_\_\_\_\_
4. Number of enrolled students who attended \_\_\_\_\_
5. Number of students making up this lab from other sections \_\_\_\_\_
6. Did you have enough time to complete required material?        Yes                    No
  - a. If not, what material was not covered?
7. Did you have technical difficulty with the clickers?        Yes                    No
  - a. If so, what happened?
8. Any other comments about lab? Anything unusual happen?
9. Don't forget to upload your saved results file (use naming format: ss-mm-dd.csv)!

---

## References

- Beatty, I. D. (2004), "Transforming Student Learning With Classroom Communication Systems," Educause Center For Applied Research Research Bulletin [online]. Available at <http://net.educause.edu/ir/library/pdf/ERB0403.pdf>.
- Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006), "Designing Effective Questions for Classroom Response System Technology," *American Journal of Physics*, 74, 31–39.
- Caldwell, J. E. (2007), "Clickers in the Large Classroom: Current Research and Best-Practice Tips," *CBE Life Sciences Education*, 6, 9–20.
- Carnaghan, C. & Webb, A. (2006), "Investigating the Effects of Group Response Systems On Student Satisfaction, Learning And Engagement In Accounting Education," *Social Science Research Network* [online]. Available at <http://ssrn.com/abstract=959370>.
- Cook, T. D. (2002), "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them," *Educational Evaluation and Policy Analysis* 24, 3 175–199.
- delMas, R., Garfield, J., Chance, B., & 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, California.

- Duncan, D. (2005), *Clickers in the Classroom: How to Enhance Science Teaching Using Classroom Response Systems*, San Francisco, CA: Pearson.
- Garfield, J. (2006), "Collaboration in Statistics Education Research: Stories, Reflections, and Lessons Learned," in *International Statistical Institute Proceedings of the Seventh International Conference on Teaching Statistics* [online]. Available at [http://www.stat.auckland.ac.nz/~iase/publications/17/PL2\\_GARF.pdf](http://www.stat.auckland.ac.nz/~iase/publications/17/PL2_GARF.pdf).
- Howe, K. R. (2004), "A Critique of Experimentalism," *Qualitative Inquiry* 10 1, 42–61.
- Kirkpatrick, D. L. & Kirkpatrick, J. D. (2006), *Evaluating Training Programs: The Four Levels*, San Francisco, CA: Berrett-Koehler.
- Light, R. J., Singer, J. D., & Willett, J. B. (1990), *By Design: Planning Research on Higher Education*, Cambridge, MA: Harvard.
- McGowan, H. M. (2009), *Experimentation Methodologies for Educational Research with an Emphasis on the Teaching of Statistics*, unpublished doctoral dissertation.
- McGowan, H. M. & Gunderson, B. K. (2010), "A Randomized Experiment Exploring How Certain Features of Clicker Use Effect Undergraduate Students' Engagement and Learning in Statistics, *Technology Innovations in Statistics Education*," 4 [online], Available at <http://escholarship.org/uc/item/2503w2np>.
- Nunnally, J. C. (1978), *Psychometric Theory*, New York: McGraw-Hill.
- Pinheiro, J. C. & Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York, NY: Springer-Verlag, *Statistics and Computing Series*.
- Raudenbush, S. W. & Bryk, A. S. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods* 2nd edition, Newbury Park, CA: Sage.
- Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995), "The development and validation of the Survey of Attitudes Toward Statistics," *Educational and Psychological Measurement*, 55, 868–875.
- Slavin, R. E. (1984), *Research Methods in Education: A Practical Guide*, Englewood Cliffs, NJ: Prentice-Hall.
- SMER. (2007), "Using Statistics Effectively in Mathematics Education Research (SMER): A Report from a Series of Workshops Organized by the American Statistical Association." [online] Available at <http://www.amstat.org/education/pdfs/UsingStatisticsEffectivelyinMathEdResearch.pdf>.
- Wu, C. F. J. & Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: John Wiley and Sons.

Zhu, E. (2007), "Teaching with clickers," CRLT Occasional Paper Number 22 [online].  
Available at [http://www.crlt.umich.edu/publinks/CRLT\\_no22.pdf](http://www.crlt.umich.edu/publinks/CRLT_no22.pdf).

---

Herle M. McGowan  
North Carolina State University  
2311 Stinson Drive  
Campus Box 8203  
Raleigh, NC 27695-8203  
<mailto:hmmcgowa@ncsu.edu>  
Phone: 919-515-0634

---

[Volume 19 \(2011\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Resources](#) | [Editorial Board](#) |  
[Guidelines for Authors](#) | [Guidelines for Data Contributors](#) | [Guidelines for Readers/Data](#)  
[Users](#) | [Home Page](#) | [Contact JSE](#) | [ASA Publications](#)