USING AND EVALUATING RESAMPLING SIMULATIONS IN SPSS AND EXCEL*

The power of computing technology has increased at an astounding rate in the last decade. Today, the personal computer plays a key role in most introductory statistics courses, freeing students from "computational drudgery" as well as enabling a sharper instructional focus on data analysis and the interpretation of statistical results. Computers have also come to play an important role in teaching statistical concepts through simulations. Despite the increased popularity of computer-based statistical simulations, there have been few empirical evaluations of their effectiveness. In this paper, I describe and evaluate three computer-assisted simulations developed for use with SPSS and Microsoft Excel. The simulations are designed to reinforce and enhance students' understanding of sampling distributions, confidence intervals, and significance tests. Results of the evaluation reveal that these simulations can help improve students' comprehension of some of the most difficult material they encounter in the introductory social statistics course.

BRAD SMITH
University of Texas at Austin

COMPUTING IN THE STATISTICS CLASSROOM

IN THE NOT-SO-DISTANT PAST, students needed to learn to compute statistics such as correlation and regression coefficients by hand because the computational power to do so was not widely available. However, as a result of the diffusion of personal computer technology and the tremendous growth in processing power over the last 15 years, statistical software applications such as SPSS and SAS are now widely used in introductory undergraduate statistics courses, freeing students from "computational drudgery" so that they can concentrate on matters of data analysis and interpretation of statistical results (Laviolette 1994). In turn, this trend has encouraged the ongoing paradigm shift among statistics educators from a formula-based approach toward a more conceptual, hands-on pedagogy. In a recent report on statistics education, the Mathematical Association of America called for "more data and concepts; less theory, fewer recipes" (Moore 1999:*xiv*). In addition to its role in automating computations, the computer is also an important instructional tool for teaching students how to analyze data. Many methods and statistics textbooks (e.g., Babbie 2003; Schutt 2001) now incorporate activities employing sample social science datasets and/or student versions of data analysis software such as SPSS. In learning to analyze data and interpret results, students gain skills useful both in the marketplace and in graduate school.

Yet another way in which the personal computer plays an important role in the statistics classroom is through simulations. Computer-based simulations differ from the

*Feedback provided by Sheryl Smith, Jean H. Shin, and by several anonymous reviewers proved beneficial in the process of writing and revising. Sample datasets, spreadsheets and student handouts are available upon request from the author. Please address all correspondence to the author at Population Research Center, The University of Texas at Austin, 1 University Station G1800, Austin, Texas 78712-0543; e-mail: bsmith@prc.utexas.edu.

Editor's note: The reviewers were, in alphabetical order, Norah Peters-Davis, Linda Marie Fritschner, Judith Stull, and Allison L. Vetter.

computer uses outlined above because they are typically used to teach concepts rather than to automate computations or teach students how to analyze data. For example, a computer simulation might display an animation demonstrating the cause-and-effect relationship between sample size and the standard deviation of a sampling distribution. Several recent introductory statistics textbooks (Hurlburt 1998; Moore 1999) come with CDs containing at least a few computer simulations. Other texts (Doane, Mathieson, and Tracy 2001; Velleman 2002) revolve around a series of computer simulations. Dozens of statistical simulations are available for public use on the World Wide Web (Lane 2000; West n.d.).

Despite the widespread availability of these tools, their instructional effectiveness has rarely been assessed. A recent review of the literature found 48 journal articles in disciplines including statistics, sociology, and psychology that recommended the instructional use of computer simulations (Mills 2002). However, only two articles actually evaluated the impact of simulations on student outcomes. The first evaluation was conducted by Weir, McManus and Kiely (1990), who designed simulations to teach undergraduates about several concepts, including the standard error of the mean and the F distribution. Their simulation allowed students to vary input parameters and see the resulting changes in the outcome of the experiment. For example, the simulation illustrating the standard error of the mean allowed students to change the sample size and monitor changes in the statistic. The researchers assessed the effectiveness of their simulations by including open-ended questions on a course assessment, concluding that the simulation was particularly beneficial for students of lower ability (based on previous grades). The second evaluation was conducted more recently by delMas, Garfield, and Chance (1999), who evaluated the benefits of a computer simulation where students drew multiple samples from any of a variety of population distributions in order to observe the long-run behavior of the mean and standard deviation of the sampling distribution. Their evaluation of the simulation suggests that it provided an effective supplement to book and lecture-based methods of instruction: the percentage of students who answered correctly, or at least used good reasoning in selecting an answer, increased from 16 percent before the exercise to 72 percent after the exercise.

While these results suggest that simulations can be beneficial to statistics students, more research is clearly needed to explore the breadth and limits of their utility in the classroom. In particular, we need to investigate whether simulations using software already employed in many classrooms and computer labs (e.g., SPSS and Excel) can be adapted to provide the same benefits as the special-purpose simulation software that has already been evaluated. Below, I discuss the general benefits that simulations offer, highlight three main types of computer simulations that have been employed in the past, then describe and evaluate a series of three new SPSS-based computer simulations designed to improve student understanding of some of the most difficult material in the introductory undergraduate statistics course.

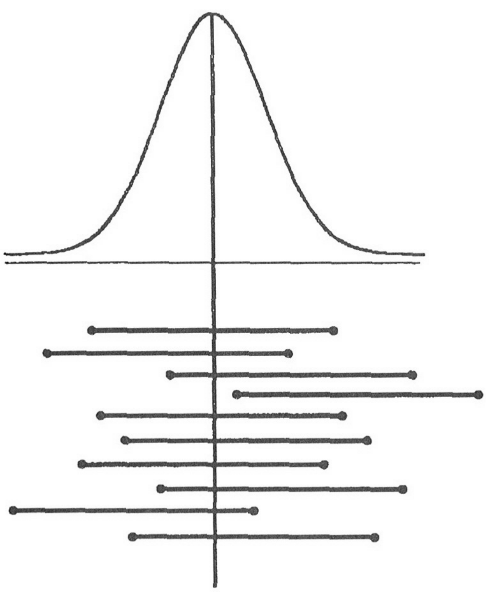
BENEFITS OF SIMULATIONS

Computer simulations provide several benefits to students learning statistics. One important benefit is that computers enable illustrations of concepts that would be difficult or impossible to demonstrate with handouts or chalkboard diagrams. For example, Moore (1999:302) uses the chart reproduced in Figure 1 to illustrate the idea behind a 90 percent confidence interval. The curve at the top represents the distribution of a continuous variable, X, at the population level. The vertical line down the center of the graph marks the population mean, μ. Each horizontal line represents the results from one round of a simulation: a sample was drawn, and a 90 percent confidence interval was computed using the formula $\bar{x} \pm z^{*}\frac{\sigma}{\sqrt{N}}$.

The leftmost point of each line represents the lower bound of the confidence interval and the rightmost point is the upper bound. Students who can read the description and imagine the process that created the diagram find this figure to be a useful way to understand confidence intervals. They can see that a 90 percent confidence interval returns the results one would expect from its definition: nine out of ten, or 90 percent, of the horizontal lines cross the line representing the population mean. Other students who come to statistics with less mathematical experience or who are not visual learners may find it difficult to interpret what the different components of the graphic represent. For many students, such displays are more likely to obscure than to illustrate the concept being taught (Trumbo 1994).

Rather than confronting students with a two-dimensional representation of a simulation, a supplementary approach involves students in the process which created the diagram. Such a process might begin with students using the computer to draw ten random samples from a large dataset (simulating the population) where the mean is already known, then computing a 90 percent confidence interval for each sample. After computing their confidence intervals, students would be able to create their own graph emulating the one in Figure 1 and see how many of their confidence intervals actually captured the population mean. Similar activities requiring students to draw repeated random samples from a population would be virtually impossible to conduct in any timely fashion without computer assistance (Trumbo 1994).

Figure 1. Graphical Display of 90 Percent Confidence Intervals



Adapted from Moore 1999

Computer simulations can also be an important part of programs to transform students from passive receivers of communicated knowledge into active participants in gaining statistical prowess (Cobb 1993; Keeler and Steinhorst 2002; Mills 2002; Schacht and Stewart 1992). In a course where many students feel as if the concepts are described in a foreign language (Mathieson, Doane, and Tracy 1995), simulations can help students feel comfortable by allowing them to interact directly with statistical principles. They can see, for example, how changing the a value in a significance-testing simulation changes the rate at which the wrong decision is made to reject the null when it is actually true. As a result, selecting an α value does not seem like pulling a number out of the air but rather represents the setting of a tolerance for risk.

THREE TYPES OF SIMULATIONS

Previous uses of computer-based statistical simulations fall into three categories: 1) stand-alone simulation software, 2) simulations utilizing existing personal computer or mainframe software applications such as Stata or Minitab, and 3) World Wide Web (WWW)-based simulations. Below, I describe each type of simulation and some representative examples. Readers interested in a comprehensive guide to statistical simulations should consult Mills (2002).

Standalone Simulation Software

Standalone simulation applications range from the commercially available and comprehensive Resampling Stats package (Simon and Bruce 2002) to smaller instructor-created, special-purpose routines such as Dimitrova, Maisel and Persell's (1993) ISEE program or Halley's (1991) GENSTAT program. While the content of each application differs, each allows students to experiment with data to see statistical principles illustrated dynamically. Many simulations in this group use resampling techniques, allowing students to draw repeated samples from a large dataset. Because students can query the entire dataset to learn the true "population" parameters, they are able to compare their results from each sample to the "population" to make conclusions about the accuracy of sampling or techniques of statistical inference.

The key strength of standalone simulation applications, particularly commercial packages, is that they require less work on the part of the instructor than programming an original simulation. Since the simulations are prepackaged, instructors need only ensure that the software is compatible with existing computer resources and provide guidance to students as they work through the simulation. However, commercial simulation packages can be costly to install and maintain on campus networks. Smaller-scale standalone applications that instructors write and share with others are less expensive but also less versatile—for example, GENSTAT and ISEE only run on a PC-compatible platform and not on Macintosh machines. They may also become difficult to maintain as computer technology advances (many older DOS and Windows applications are useless under the most recent version of Windows). Both kinds of standalone applications also share an interface liability. In courses where students are exposed to some type of data analysis software such as SPSS or Stata, adding a second (or third) software interface to the mix has the potential to further alienate students with weak computer skills. Investing class time in teaching students how to use simulation software is also unlikely to produce the same marketable skills in students as providing instruction using common commercial software packages.

Simulations Within Existing Software

A second group of simulations harnesses existing software tools to teach statistical concepts. Because the design of both Minitab and Stata makes it easy to implement Monte Carlo procedures, several simulation routines have been created for these environments. To illustrate the idea of confidence intervals, Kennedy, Olinsky, and Schumaker (1990) designed a simulation to be executed in Minitab. Students begin by drawing a sample from a dataset (the "population") provided by the instructor. By aggregating results with classmates, students can measure what proportion of the time their intervals actually capture the true population mean. Working in Stata, Ferrall (1995) created a set of programs that illustrate the Central Limit Theorem and the impact of non-normality of the error term on sample regression coefficients. Simulations that run within existing statistical software are advantageous because they reduce the cost of using separate software for instruction in data analysis and for teaching statistical principles. Unfortunately, unless students are required to own a computer and to purchase Minitab or Stata as part of course requirements, their experience with the software will likely not extend beyond the computer lab. Furthermore, instructors who favor software for which few, if any, simulations have been created (e.g. SPSS) face a difficult choice: add a second program for students to learn, or redesign the curriculum around new software.

World Wide Web (WWW)-Based Simulations

The third and newest type of simulation is not installed on personal or mainframe computers but is delivered over the Web. The advent of the Internet and the Java programming language made it possible to craft small software applications, or applets, that download over the Internet and run within a Web browser. As a result, the software is platform-independent, enabling the same simulation to run as easily on a PC as on a Macintosh or a Unix workstation. One of the most comprehensive collections is Lane's (2000) Virtual Lab in Statistics which contains a collection of 21 Java applets illustrating statistical concepts ranging from the mundane (mean and median) to the advanced (two-way ANOVA). The simulation on sampling distributions, for example, allows students to execute a resampling experiment based on a student-chosen population distribution, sample size, and number of repetitions. As the experiment progresses, students can watch the sampling distribution being built sample by sample, enabling them to see that a sampling distribution for the sample mean is the aggregation of all possible sample means of a given size. They can also see that the sampling distribution is approximately normal for large sample sizes regardless of the shape of the population distribution. Marden (2000) has also developed a comprehensive set of instructional simulations. One applet allows students to practice assessing the magnitude of a relationship by examining scatterplots. Students are presented with a group of four unlabeled scatterplots and then are asked to match each plot to the correct correlation coefficient chosen from a list of four possibilities. After matching all four plots to correlation coefficients, students receive immediate feedback on the accuracy of their answers. A feature permitting a user to keep track of his or her running score and compete against classmates is particularly popular with students.

Simulation applets cover nearly every topic discussed in an introductory undergraduate statistics course. Because they are easily accessible to students from any Internet-connected computer, they also allow students to study and review concepts outside of the computer lab. Their notable disadvantage is that in most cases, these simulations tend to involve the student mostly as a spectator. Students click a button on the screen and the animated simulation takes place before their eyes. In the best WWW-based simulations, the graphics do a good job of communicating the process being simulated. For example, Lane's (2000) sampling distribution simulation uses small bars dropping onto a histogram to represent individual observations within a sample. Just below this histogram, a bar representing the mean of that sample drops onto another histogram representing the sampling distribution. However, in other simulations, graphics are either absent or serve as poor translators of the process behind the simulation. As a result, WWW-based simulations are most effective when chosen carefully and accompanied by teaching techniques that induce students to think more deeply about the demonstrations they witness.

THREE NEW RESAMPLING SIMULATIONS

I created simulations to elucidate three concepts (sampling distributions, confidence intervals and significance tests) that are particularly problematic for undergraduate students. Despite the array of simulations available, I chose to design new ones because I was not able to find any written for the software I used in my courses (SPSS and Excel). Some critics argue that bypassing Web-based simulations that illustrate the same concepts and require no special software in favor of SPSS and Microsoft Excel adds an additional and unnecessary layer of complexity to the instructional process. However, using these common software tools provides benefits unavailable through purely Web-based simulations. First, the additional exposure to these common commercial software packages is more likely than special-purpose simulation software to confer to students a marketable skill advantage. Furthermore, asking students to use and move data between SPSS and Microsoft Excel to complete the simulation highlights the process of drawing repeated random samples in a way that a completely animated demonstration on a WWW-based simulation cannot.

The simulations use resampling techniques—the process of taking repeated random samples from a fixed and known population—because of the way that resampling demystifies statistical inference (Good 2001; Simon 1997). In real-world applications of inference, parameters such as the population mean or the population regression coefficient are unknown, so the analyst never knows exactly how accurate sample estimates are. In resampling, however, repeated random samples are drawn from a hypothetical "population" where the parameters (mean, proportion, or even a regression coefficient) are known. In the simulations described below, each "population" consists of data I either generate or adapt from a secondary dataset such as the General Social Survey. Because in this unrealistic but instructionally useful case the truth is known about the population parameters, students immediately see how good (or bad) their estimates from each new sample are. Experience with estimating known population parameters also helps clarify the distinctions between population and sample and between parameter and estimate.

These simulations are designed to be used in a computer lab with SPSS and Microsoft Excel and are tailored to complement the text (Moore 1999) I use in my course, but could easily be adapted for use in any introductory statistics course and with any statistical software package that allows random selection of observations in the active data set. I chose SPSS simply because of its widespread availability on campus and because its random sampling capability is easily accessible from the menus, meaning students do not have to learn programming syntax (although teaching some syntax or providing pre-written programs can speed the process significantly). Each of the three simulations described below involves the same basic procedure adapted to different concepts. Students begin by taking a series of simple random samples from a dataset which serves as a "population." They then complete a hands-on computer activity, participate in a group discussion led by the lab instructor, and answer follow-up questions for course credit. The questions are designed to cause students to think more deeply about the key objectives of the exercise by asking them to compare their own results to the predictions of statistical theory. Datasets, student handouts, and spreadsheets are available from the author.

Sampling Distribution for the Sample Mean

We begin the section of the course that deals with statistical inference with a discussion of the difference between the population and a sample. Students seem to have little difficulty with this distinction. However, many students stumble when they encounter the sampling distribution concept. It is not difficult to imagine drawing one sample and computing one sample mean, but the idea of taking every possible sample of a given sample size seems much harder to grasp for students without much mathematical experience (Schwarz and Sutherland 1997). I have three objectives in the lesson on sampling distributions. First, I want students to understand that a sampling distribution represents the collection of all possible sample estimates for a given sample size. Second, I want students to know what it means if a statistic is an *unbiased* estimator of a population parameter. Third, I want students to understand the relationship between efficiency and sample size. The following exercise was designed to communicate these three ideas to students.

Students begin the exercise by opening an SPSS data file containing data from our "population" of about 6,000 cases representing the family income for each of the 6,000 freshmen coming to the local state university next fall. I explain to students that we will treat these data as the population and take many samples from the population to investigate the performance of our sample estimate ($\bar{x}$) of the population mean (μ). Students first identify the population mean by applying SPSS's descriptive statistics procedure to the entire dataset. After identifying the population parameter they will be estimating, students proceed to draw 10 samples at each of three sample sizes. I have students select 10 samples where N=60, 10 where N=600, and 10 more where N=3,000 (the precise sample sizes are unimportant, as long as there are substantial increases from one stage to the next). I tell students we will investigate how the accuracy of our estimates changes as sample size increases. After they complete the resampling procedure, I ask students to enter their 30 sample means into SPSS and compute the mean and standard deviation of their ten sample means for each of the three sample sizes.

To illustrate the three key ideas I am trying to communicate in class, I collect each student's 30 sample means (about 1,000 total in my classes of 36 students) via email and tell them I am going to aggregate their data in a new SPSS file.[1] I close the simulation by leading students through an analysis of the dataset containing the more than 300 sample means for each of the three sample sizes:

- First, I ask students to use SPSS to create a histogram of the sample means for each sample size. I then define a sampling distribution by explaining that the distribution of our sample means is a rough approximation of the sampling distribution for $\bar{x}$ at any given sample size.

- Second, I illustrate the idea of an unbiased estimator by showing that the overall mean of the 300 individual sample means that students obtained for each sample size is very close to the population mean, μ, and that the standard deviation is very close to

$$\frac{\sigma}{\sqrt{N}}.$$

- Third, I demonstrate how increases in sample size improve statistical efficiency by showing that as sample size increases, the standard deviation of our sampling distribution decreases.

As a response to this experiment, I ask students to explain in short answer format what a sampling distribution is, what "unbiased" means and how efficiency is related to sample size.

After discussing and experimenting with sampling distributions, the course proceeds to common techniques of inference that students are likely to encounter in data analysis. In this portion of the course, the greatest struggle is pushing students beyond rote application of a series of steps to an in-depth understanding of what their computations really mean. I designed two simulations to deepen students' understanding of confidence intervals and significance tests. Because these two simulations essentially apply the same resampling framework to new concepts, they are described in less detail.

Statistical Confidence

One common explanation for a C% confidence interval is that there is a (C/100) probability that the interval will contain the population parameter (Frankfort-Nachmias 1997:522). Faced with only one sample and only one confidence interval, however, students often have difficulty comprehending what this explanation really means. The following exercise allows students to compute confidence intervals on multiple samples and investigate how often their intervals actually do contain the population parameter. My instructional objective getting students to understand that the level of confidence (e.g., 95% or 0.95) in a confidence interval represents the likelihood that their

[1] In order to keep students focused on the concepts at hand, I typically use a collection of means generated by students from a previous semester rather than taking the time to compile students' sample means during the class period. Students' responses in class suggest that they do not feel misled by this substitution.

confidence interval captures or contains the parameter of interest.

Students begin with a SPSS data file consisting of a single continuous variable with approximately 1,000 observations. I ask them to use SPSS to compute the mean for our hypothetical population (again, the "population" here is just all the cases in the active data set). Next, using SPSS's random sampling feature, they take 20 samples at a moderate sample size (e.g., N = 100) and compute the sample mean for each. Then, I ask students to open a Microsoft Excel spreadsheet that I provide to them. Before the class meeting, I set up the spreadsheet to automate the computation of their 20 confidence intervals. After students enter the population mean and their 20 sample means, the spreadsheet computes the bounds of each confidence interval, assesses whether each interval contains the population mean, and calculates the proportion of times the confidence interval is "right" (contains the population parameter). I ask students to verify the confidence interval and decision by hand for at least one sample mean. After students are sure that the spreadsheet is making the right decision, students can vary the level of confidence, C, to see how the rate of capturing the population mean changes as C changes. As a response to the exercise, I ask students to explain in a brief paragraph the relationship between the confidence level, C, and the likelihood that their confidence intervals will contain the population mean.

Alpha Levels in Significance Tests

Most students are able to follow the basic procedures for testing hypotheses. However, even after discussing what Type I and II errors signify, students inevitably ask, "But where do you get alpha (α)?" Even after discussing this idea during at least two class periods, it was apparent that students did not fully grasp the underlying concepts, so I designed an exercise that illustrates how changing α, the significance threshold, affects the risk of making the wrong decision—to reject the null hypothesis when it is actually true. As in the previous simulations, students begin with a hypothetical population. In this case, I present students with a data set containing a single continuous variable with about 1,000 cases. Students establish the mean of the hypothetical population by using SPSS's descriptive statistics function and take 30 samples at a moderate sample size, recording the mean and standard deviation of each sample. They then enter their sample means and sample size in a Microsoft Excel spreadsheet. The spreadsheet computes a test statistic for the null hypothesis H_0: $\mu = k$, where k is the mean from our hypothetical population for each sample and also automates the decision about the null hypothesis, telling students whether or not the null hypothesis was rejected. After students verify the accuracy of the spreadsheet's work by hand, they compute the frequency with which the null hypothesis (known to be true in this case) was rejected in error. Students can increase and decrease the α level on the spreadsheet to see that the rate of rejecting a true null hypothesis varies directly with α. To close the exercise, I emphasize that one way to think about α is as a measure of acceptable risk—the risk you are willing to take in rejecting the null hypothesis when it is actually true. To wrap up, I ask students to write a brief paragraph explaining how our results illustrate what α represents in significance testing.

EVALUATION OF THE SIMULATIONS

In two recent sections of my undergraduate statistics course, I undertook evaluations of each of these simulations. In the first case, I taught a small (12 students) 3-week summer session course at a major Midwestern university. Because of time constraints, I was able to evaluate only the sampling distribution and significance testing simulations. Recently, I taught a larger section (32 students) of the course at a small liberal arts college and evaluated all three simulations. In both cases, the general approach to

Table 1. Results from Evaluation of Sampling Distributions Exercise: Proportions of Correct and Partially Correct Answers to Quiz Questions in Experimental (N=22) and Control (N=21) Groups

	Partially Correct Answers		*Correct Answers*		
Question	**Experimental**	**Control**	**Experimental**		**Control**
What is a sampling distribution?	0.18	0.14	0.27	*	0.05
What does it mean when we say a sampling distribution is unbiased?	0.41	0.24	0.50	*	0.05
What is meant by statistical efficency?	0.27	0.24	0.73	*	0.38

Note: * = one-tailed, two-sample difference of proportions test significant with $p < 0.05$.

evaluation was the same. I randomly assigned students either to an experimental or a control group. When we reached the appropriate point in the semester, I gave my standard lecture on the topic at hand. The following day, I took the experimental group of students into the computer lab and worked through the appropriate simulation. No new material was introduced in the lab sessions. The control group was given the afternoon off. The next class period, I gave a short quiz (closed book and notes) or added a question to a pending course examination.[2] The assessments required short written answers and were tailored to gauge the degree to which students understood the objectives of each simulation. The standards for assessing correct answers were developed from the objectives that the simulations were designed to achieve. I graded students' answers without knowledge of who was in the control and experimental groups, assigning 0 points if the answer was incorrect, 1 point if the answer was partially correct, and 2 points if the answer was complete and correct. After each assessment, I took the control group through the same simulation that the experimental group had experienced.

How did the experimental group fare compared to their colleagues in the control group? To evaluate the effectiveness of the sampling distributions simulation, I used a three-question, open-ended assessment. Each of the questions was tailored to one of the objectives I had emphasized both in lecture and in the computer lab. I asked students to 1) define a sampling distribution, 2) explain what it means that a statistic is an unbiased estimator of a population parameter, and 3) explain the relationship between efficiency and sample size. Correct answers for the first question were those which explained that a sampling distribution is the collection of all samples of a given sample size. For the question on bias, I marked as correct answers which indicated that to be unbiased, a sampling distribution had to be centered on a population parameter. For the last question, students who answered correctly were able to explain the idea that as sample size goes up, the spread of a sampling distribution goes down.

In Table 1, I present the proportions of students in the control and experimental groups who received partially correct and completely correct marks for each question. To ascertain whether the differences observed were statistically significant in the expected direction, I used a one-tailed two-sample proportions test.[3] Comparing the rate of partially and completely correct answers between the two groups in the experi-

[2]I did not count the grades from either type of assessment used in the evaluation of the simulations toward a student's overall course grade.

[3]The results of any statistical inference on a sample this small should be interpreted with caution. Because of small sub-sample sizes, the assumptions for a two sample proportions test ($np \geq 5$ and $n(1-p) \geq 5$) are occasionally not met. Fortunately, the group differences which are flagged by the test as statistically significant are also large enough to be substantively meaningful.

Table 2. Results from Evaluation of Significance Test Exercise: Proportions of Correct and Partially Correct Answers to Quiz Questions in Experimental (N=22) and Control (N=21) Groups

	Partially Correct Answers		*Correct Answers*		
Question	**Experimental**	**Control**	**Experimental**		**Control**
What does the α value mean in a significance test?	0.43	0.29	0.43	*	0.29

Note: * = one-tailed, two-sample difference of proportions test significant with $p < 0.05$.

ment revealed some interesting differences. For all three objectives, the proportion of students who gave partially correct answers in the experimental group was larger than but not statistically distinguishable from the proportion in the control group. In the case of completely correct answers, statistically significant differences between the control and experimental groups were noted across all three questions. Not only was the difference statistically significant, the difference in scores of the experimental and control groups was also substantively large. The rate of correct answers was from two to ten times as large in the experimental as in the control group.

The second exercise was designed to improve students' understanding of α, the significance threshold. Students were asked to explain what the α value means in significance testing. To be deemed correct, an answer had to both mention and explain the idea of "acceptable risk"—that α represents the proportion of the time we are willing to make an incorrect decision by rejecting the null hypothesis when it is actually true. Although the experimental group did indeed provide both partially correct and completely correct answers at a substantially higher rate than the control group, with the relatively small sample sizes employed here the difference was not statistically significant.

For the exercise on statistical confidence, I measured student comprehension with the question, "What does 'confidence' mean in the phrase 'confidence interval'?" Correct answers were those which recognized and explained that "confidence" represents how sure we are that our interval captures the population parameter of interest. On this question, the results were mixed. Surprisingly, for partially correct answers, students in the control group actually did better than the experimental group. The difference was not, however, statistically significant. Regarding completely correct answers, the evidence strongly suggests that experimental group members better understood the concept than did control group members. Almost half of experimental group members answered the question correctly, but no control group member wrote a completely correct answer.

DISCUSSION

Each of these three exercises was designed to elucidate an important concept in statistical inference through the same basic structure of resampling—students take a series of

Table 3. Results from Evaluation of Confidence Intervals Exercise: Proportions of Correct and Partially Correct Answers to Quiz Questions in Experimental (N=15) and Control (N=11) Groups

	Partially Correct Answers		*Correct Answers*		
Question	**Experimental**	**Control**	**Experimental**		**Control**
What does "confidence" mean in the phrase "confidence interval"?	0.47	0.73	0.47	*	0

Note: * = one-tailed, two-sample difference of proportions test significant with $p < 0.05$.

samples from a hypothetical population in order to investigate how a particular procedure or statistic performs across many samples. Because each of the exercises involves a comparison between sample estimates and population parameters, these simulations also serve to increase students' exposure to the concept of a sampling distribution, which is both a fundamental feature of classical statistical inference as well as a significant stumbling block for students encountering statistics for the first time.

The results of the evaluation generally bear out the utility of these resampling exercises in teaching the fundamentals of statistical inference. Of the ten possible outcomes (five questions, comparing both partially and completely correct answers) where differences in comprehension between the experimental and control groups were assessed, the experimental group outperformed the control group in nine cases. The differences in performance between the experimental and control groups were statistically significant in four of the nine cases.

In addition to the measurable improvement in student understanding conferred by these simulations, my classroom experience suggests that they are useful in other ways as well. Laboratory exercises such as these can, for example, pique students' interest in statistics just as they encounter some of the most difficult material in the course. It has been my experience that an interesting lab exercise can re-energize students much more effectively than the best lecture I can deliver. These techniques are also useful because of their flexibility. The simulations discussed here are not tied to dedicated software but can be executed in any statistical software package that permits a user to sample from an existing dataset. Variations on these exercises could easily be created to teach about other statistical principles or tests, such as the Central Limit Theorem.

While I am enthusiastic about the improvements these exercises made in my classroom, simulations still offer only supplementary instructional strategies. The evaluation I conducted, as well as the evaluations conducted by Weir, McManus and Kiely (1990) and delMas, Garfield and Chance (1999), all presume the foundation of an organized presentation in class and a well-written textbook. Simulation methods such as those described here are not a replacement for these methods of instruction, but an effective alternative method of communication with students who may be spatial or kinesthetic rather than linguistic learners. As the effort to ascertain the effectiveness of computer simulations continues, additional research is needed to more fully explore the benefits of other types of simulations as well as the range of topics or types of students for which simulations are likely to be particularly beneficial.

REFERENCES

Babbie, Earl R. 2003. *The Practice of Social Research.* Belmont, CA: Wadsworth Publishing Company.

Cobb, George W. 1993. "Reconsidering Statistics Education: A National Science Foundation Conference." *Journal of Statistics Education* 1(1). Retrieved April 11, 2001 (http://www.amstat.org/publications/jse/v1n1/cobb.html).

delMas, Robert C., Joan Garfield, and Beth L. Chance. 1999. "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning." *Journal of Statistics Education* 7(3). Retrieved May 10, 2002 (http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm).

Dimitrova, Ganka, Richard Maisel and Caroline Hodges Persell. 1993. "Using and Evaluating ISEE, A New Computer Program for Teaching Sampling and Statistical Inference." *Teaching Sociology* 21:341-51.

Doane, David P., Kieran D. Mathieson, and Ronald L. Tracy. 2001. *Visual Statistics 2.0.* New York: McGraw Hill.

Ferrall, Christopher. 1995. "Interactive Statistics Tutorials in Stata." *Journal of Statistics Education* 3(3). Retrieved May 10, 2002 (http://www.amstat.org/publications/jse/v3n3/ferrall.html).

Frankfort-Nachmias, Chava. 1997. *Social Statistics for a Diverse Society.* Thousand Oaks, CA: Pine Forge Press.

Good, Phillip. 2001. *Resampling Methods.* Bos-

ton, MA: Birkhauser.
Halley, Fred S. 1991. "Teaching Social Statistics with Simulated Data." *Teaching Sociology* 19:518-25.
Hurlburt, Russell T. 1998. *Comprehending Behavioral Statistics*. 2d ed. Pacific Grove, CA: Brooks/Cole.
Keeler, Carolyn M. and R. Kirk Steinhorst. 2002. "Using Small Groups to Promote Active Learning in the Introductory Statistics Course: A Report from the Field." *Journal of Statistics Education* 3(2). Retrieved May 17, 2002 (http://www.amstat.org/publications/jse/v3n2/keeler.html).
Kennedy, Kris, Alan Olinsky, and Phyllis Schumacher. 1990. "Using Simulation as an Integrated Teaching Tool in the Mathematics Classroom." *Education* 111:275-96.
Lane, David. 2000. "Rice Virtual Lab in Statistics." Houston, TX: Rice University. Retrieved May 28, 2002 (http://www.ruf.rice.edu/~lane/rvls.html).
Laviolette, Michael. 1994. "Linear Regression: The Computer as a Teaching Tool" *Journal of Statistics Education* 2(2). Retrieved May 17, 2002 (http://www.stat.ncsu.edu/info/jse/v2n2/laviolette.html).
Marden, John. 2000. "The CUWU Statistics Program." Champagne, IL: University of Illinois at Urbana-Champaign. Retrieved May 22, 2002 (http://www.stat/uiuc.edu/~stat100/cuwu/).
Mathieson, Kieran, David P. Doane and Ronald L. Tracy. 1995. "A Program for Visualizing Comparisons Between Two Normal Distributions." *Journal of Statistics Education* 3(1). Retrieved May 17, 2002 (http://www.amstat.org/publications/jse/v3n1/mathieson.html).
Mills, Jamie D. 2002. "Using Computer Simulation Methods to Teach Statistics: A Review of the Literature." *Journal of Statistics Education* 10(1). Retrieved May 17, 2002 (http://www.amstat.org/publications/jse/v10n1/mills.html).
Moore, David S. 1999. *The Basic Practice of Statistics*. 2d ed. New York: W.H. Freeman:.
Schacht, Steven P. and Brad J. Stewart. 1992. "Interactive/User-Friendly Gimmicks for Teaching Statistics." *Teaching Sociology* 20:329-32.
Schutt, Russell. 2001. *Investigating the Social World*. Thousand Oaks, CA: Pine Forge Press.
Schwarz, Carl James and Jason Sutherland. 1997. "An On-Line Workshop Using a Simple Capture-Recapture Experiment to Illustrate the Concepts of a Sampling Distribution." *Journal of Statistics Education* 5(1). Retrieved May 17, 2002 (http://www.amstat.org/publications/jse/v5n1/schwarz.html).
Simon, Julian and Peter Bruce. 2002. *Resampling Stats 5.0*. Arlington, VA: Resampling Stats, Incorporated.
Simon, Julian L. 1997. "Resampling Statistics." Retrieved April 23, 2002 (http://www.resample.com/content/text/index.shtml).
Trumbo, Bruce E. 1994. "Some Demonstration Programs for Use in Teaching Elementary Probability: Parts 1 and 2." *Journal of Statistics Education* 2(2). Retrieved April 11, 2001 (http://www.amstat.org/publications/jse/v3n2/trumbo.html).
Velleman, Paul. 2002. *ActivStats: 2002-2003 Release*. Boston, MA: Addison Wesley Longman.
Weir, Catherine Grant, I. Chris McManus and B. Kiely. 1990. "Evaluation of the Teaching of Statistical Concepts by Interactive Experience with Monte Carlo Simulations." *British Journal of Educational Psychology* 61:240-7.
West, R. Webster. N.d. "Java Applets." Retrieved April 22, 2002 (http://www.stat.sc.edu/~west/javahtml).

Brad Smith received his Ph.D. from Indiana University. He has taught courses in research methods, statistics, complex organizations, and introductory sociology at McDaniel College, formerly known as Western Maryland College. Currently a postdoctoral research associate at the Population Research Center at The University of Texas at Austin, his research interests include religion, social networks, and mental health.