9. OVERVIEW OF CONSTATS AND THE CONSTATS ASSESSMENT

Steve Cohen and Richard A. Chechile
Tufts University

INTRODUCTION

ConStatS has been in development at the Tufts University Curricular Software Studio for the past nine years. From the beginning, the goal of the project was to develop software that offered students a chance to actively experiment with concepts taught in introductory statistics courses. It is a joint product of faculty from engineering, psychology, sociology, biology, economics, and philosophy. During the past nine years, there have been periods alternatively devoted to development, assessment, and classroom use.

ConStatS consists of 12 *Microsoft Windows*-based programs, grouped into five distinct parts as described below.

1. **Representing Data:** Different ways in which aggregates of data are represented in statistics, both graphically and numerically
   - *Displaying Data* − univariate data given in tables displayed in histograms, cumulative frequency displays, observed sequence graphs, and bar charts, as an initial step in data analysis.
   - *Descriptive Statistics* − univariate summary statistics describing the *center* (e.g., the mean and median), the *spread* (e.g., the variance, standard deviation, and interquartile range), and the *shape* of data.
   - *Transforming Data* − linear transformations, especially *Z* scores, and their effects on the center, spread, and shape of distributions of univariate data; also, frequently used nonlinear transformations for changing the shapes of distributions.
   - *Describing Bivariate Data* − scatterplots and summary statistics for bivariate data, with emphasis on the use of the least squares line, residuals from it, and the correlation coefficient in analyzing data to find relationships between variables.

2. **Probability**: Basic concepts in probability that are presupposed in advanced topics in statistics, such as sampling and inference
   - *Probability Measurement* − numerical probabilities as ratios, and consistency constraints on them, illustrated by having students assign numerical probabilities to alternatives in everyday situations.
   - *Probability Distributions* − the key properties of 14 probability distributions used in statistics, including the binomial and the normal, brought out by interactive comparisons between graphical displays of their probability density functions, their cumulative distribution functions, and pie charts.

3. Sampling: Gains and risks in using samples to reach conclusions about populations

Sampling Distributions − the variability and distribution of the values of different sample statistics for samples of different sizes drawn from populations having different (postulated) underlying probability distributions.

Sampling Errors − the risks of being misled when using sample statistics, obtained for samples of different sizes, as values for the corresponding population statistics.

A Sampling Problem − a game in which a simulated coin can be tossed repeatedly before deciding whether it is fair, or it is 55% or 60% biased in favor of heads, or 55% or 60% biased in favor of tails.

4. Inference: The basic frameworks of reasoning in which statistical evidence is used to reach a conclusion or to assess a claim

Beginning Confidence Intervals − repeated sampling used to show the relationship between the width of an interval, employed as an estimator for the population mean, and the proportion of the times it will cover this mean.

Beginning Hypothesis Testing − a step-by-step tracing of the reasoning involved in the statistical testing of claims about the mean of a single population or about the difference between the means of two populations.

5. Experiments: Experiments in which the user of ConStatS is the subject, for purposes of generating original data for use in the Representing Data programs

An Experiment in Mental Imagery − the classic Shepard-Metzler experiment in cognitive psychology involving the rotation of images, yielding as data the time taken for the subject to react versus the number of degrees through which the image is rotated.

PROGRAM DESIGN AND DESCRIPTION

Early versions of the software used a standard “point and click” graphical user interface (Cohen, Smith, Chechile, & Cook, 1994). Pull down menus were used to access datasets, exercises, experiments, and programs on different topics. Early classroom trials did not produce the kind of use and learning expected. Most students, left to their own devices, became lost when they had to make all the decisions. Focusing on a question and translating the question into choices offered by the program was a daunting task to most students. What seemed like elementary decisions to faculty who were designing the software (i.e., selecting a dataset and a variable to work with) proved difficult and intimidating to many students. Selecting, designing, and executing experiments proved even more difficult. These early trials demonstrated that most students were not comfortable designing their own learning pathways.

The current version of ConStatS uses a combination of devices to solve this problem (Cohen et al., 1994a). First, each program in the package is divided into a large number of "screens," no one of which confronts the student with more than a small number of closely related decisions. Figure 1 shows a typical ConStatS screen (sitting under the main menu from which programs are selected.)

9. OVERVIEW OF CONSTATS AND THE CONSTATS ASSESSMENT

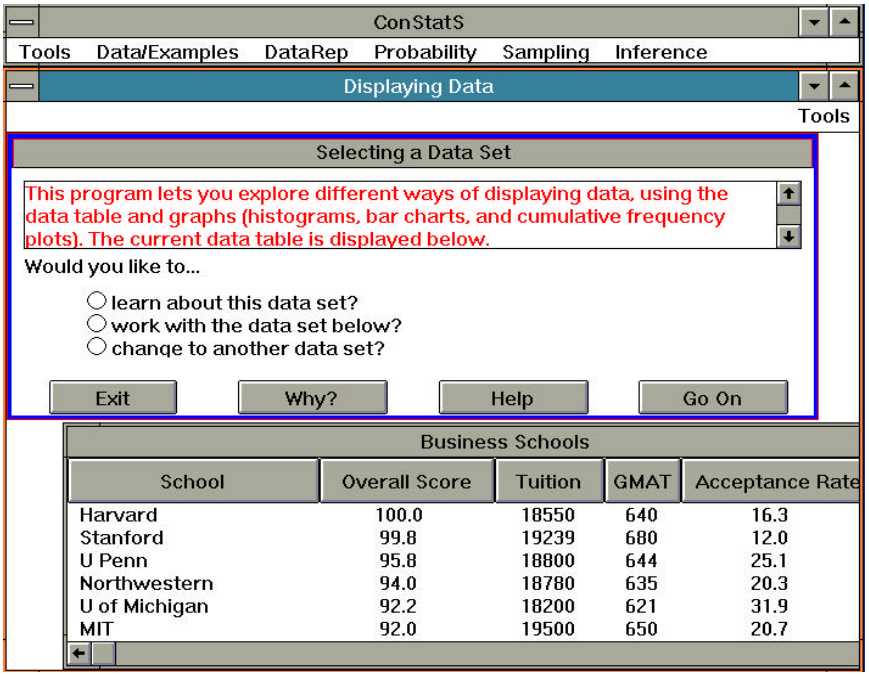

Figure 1: A typical ConStatS screen

The choices the student makes on each screen lead to different screens and pathways through the program, pathways that often loop into one another. Some screens help students prepare experiments (i.e., selecting a pathway or setting a parameter) and others are for performing experiments. Figure 2 shows a second screen from a pathway in the Displaying Data program.

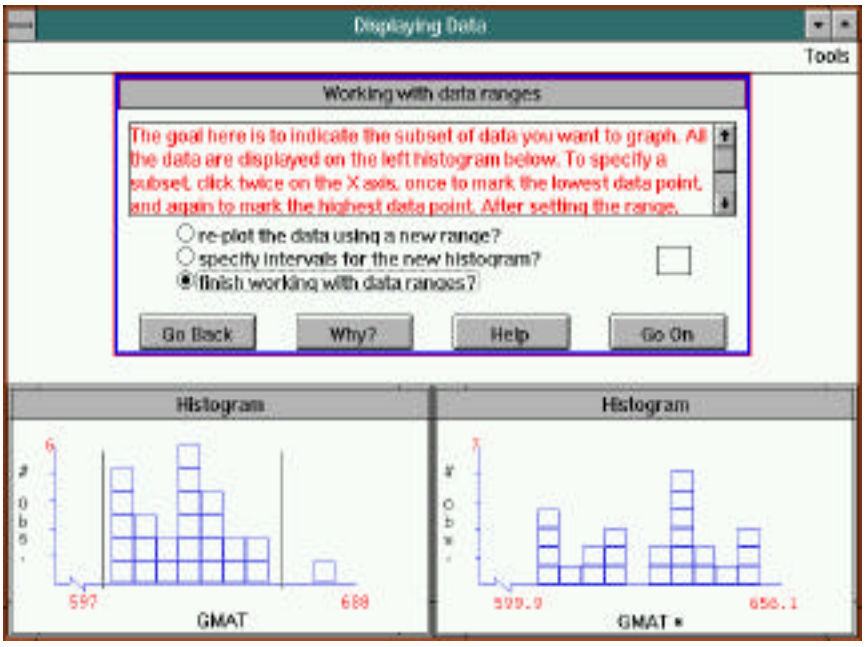

Figure 2: An experiment to examine the influence of outliers

It is an experiment screen designed to let students examine the influence of an outlier by eliminating it and seeing the resulting distribution.

The pathways provide an unobtrusive structure that helps guide the student along in an orderly fashion. More guidance is provided at some places by making the decision between a default value offered by the program (i.e., default population statistics) and a value of the student's own choosing (i.e., a user-defined population statistic). Each screen has a one or two sentence "scaffolding," which introduces the choices that have to be made. The student can always back up along a pathway to review or reconsider earlier choices.

Although the structure offers guidance and support for choosing among options, it does not interfere with students roles as active, *experimental* learners. The only questions that ever appear on the screen are ones that have to be answered to determine a desired result or to initiate a new direction. No "study questions" ever appear, nor do any other didactic elements that would tend to induce students to fall into a passive style of learning. The students are always in control, not just in the sense that they choose what to do next, but in the sense that nothing ever happens on the screen except through choices they make. Each screen presents them with a handful of choices, posed as questions. These choices are the ones that have to be made to determine what result that will appear next (e.g., the choice of data-range and the number and type of intervals in order to draw a histogram).

Finally, and most importantly, a WHY and HELP button are available on every screen, allowing access to information that will help confused students. Hitting the WHY button when facing a choice produces a reason why the choice is an appropriate one to be facing. This usually takes the form of a one sentence statement of a typical consideration that someone might focus on when making the choice. For example, hitting WHY when hesitatng over the quesion, "Do you want to change the number of intervals?" produces "Maybe the histogram will take on a very different apearance with a different number of intervals"--just the sort of thought that a good teacher might whisper in the ear of a student who is hesitating in the middle of an experiment. Hitting the HELP button produces a paragraph or two discussing the choice. This is the only place in the software where book-like elements intrude. However, even here the student is actively eliciting the information, looking for specific things that will help the student take the next step, in much the way that superior students take a quick look at one or two pages of a book when working something out in a thought experiment.

In the spirit of anchored instruction (The Cognition and Technology Group at Vanderbilt, 1990), pathways typically require students to perform a series of experiments on the same data [e.g., data on the variable High School (HS) Graduation Rates]. Early in the Displaying Data program, students select a single variable from a dataset. The pathway allows students to examine this data in experiments on using and reading histograms, cumulative frequencies, displaying subsets of data, comparing histograms and cumulative distributions, and in other experiments on univariate display. From this, we hoped that concepts would be anchored in a specific variable (i.e., an example). Similarly, when working with probability distributions, students use the same distribution (be it normal, binomial, etc.) to examine parameters, probability density, cumulative density, and so forth. At any point, students can return to the beginning of the pathway, select a different variable (or probability distribution), and move through the pathways to repeat the experiments.

Once students have worked through specific experiments and become familiar with concepts, they can turn to a facility for using the statistics and conducting data analysis. Figure 3 shows this facility from the Describing Bivariate Data program. There are no experiments, questions, or WHY and HELP buttons. Instead, students make choices from a menu as they might if they were using a data analysis package

designed for graphical user interfaces. The options available not only include topics covered in the current program, in this case bivariate regression, but also topics covered in other ConStatS programs. For instance, Transformations is included as one option among many in this screen.

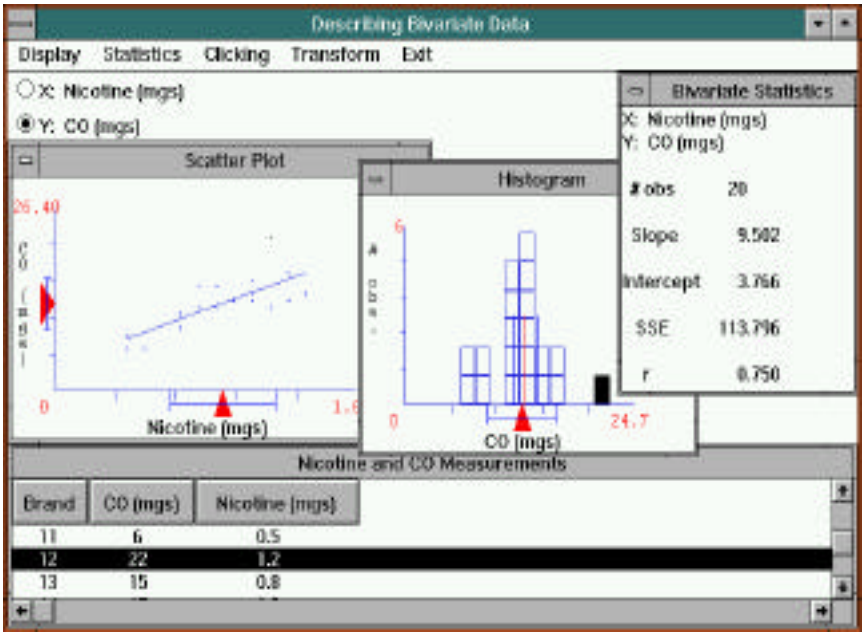

Figure 3: Using concepts examined in the Describing Bivariate Data Program

Finally, to make ConStatS useful for statistics courses taught in a variety of departments, carefully chosen datasets from several different disciplines, including psychology, sociology, economics, biology, and engineering, as well as data of general interest, were included. New datasets can readily be added by students and teachers. The emphasis of the overall package is on gaining conceptual understanding of statistics. But precisely because statistics is primarily a discipline of application, students gain such understanding best when dealing with real data that they find interesting.

ASSESSING CONSTATS

In 1991, with funding from FIPSE (Fund for the Improvement for Postsecondary Education), we began a three-year assessment of ConStatS. By that time, ConStatS had become integrated into the Tufts curriculum in several departments, including psychology, economics, and engineering. However, ConStatS, at that time, consisted of only the first 9 of the 12 programs described above.

The principal goal of the assessment was to examine learning outcomes. Several important research design decisions were made in the following areas:

Multidiscipline and multisite: We were interested in investigating whether the software was effective in a range of statistics courses, taught in a variety of departments. In addition, positive effects might be attributable to the software being used at the institution at which it was developed. To determine transferability, several outside sites were also included, with classes taught by professors uninvolved with the development of ConStatS. These classes were taught in psychology, biology, and education departments. At

least one of the outside schools was comparable to Tufts in student profile. Finally, four different classes, all at Tufts, participated as control groups. We did not try to explicitly recreate the experiments and exercises in ConStatS in the control classes (Clark, 1994). However, to help make sure the content of the control classes was similar to the content of the software, two of the control classes were taught by a member of the team that designed ConStatS.

Assumed basic skills: ConStatS was designed to teach conceptual understanding. Still, certain basic mathematical skills were assumed during the development of the software. A 10-item pretest was administered to all students, control and experimental, who participated in the assessment. The pretest included items on fractions, ratios, ordering integers, very basic algebra, and distinguishing variables from constants.

Isolating the concepts: The nine ConStatS programs used in the assessment covered hundreds of concepts. Most of the concepts, like that of an outlier (Figure 2), are covered in specific experiments in appropriate programs. However, many, if not most, concepts appear in more than one place in the software. Because the goal of assessment was to learn how effective each part of the software was in helping students acquire concepts, we needed to identify where in the nine programs each concept was encountered.

Consider the screen in Figure 4 from the Transformations program. The screen shows a step-by-step animation in progress: the data in the histogram in the lower left is undergoing a *Z*-score transformation and is replotted in the lower right. The process is illustrated for each data point, until the student presses the END STEP button.

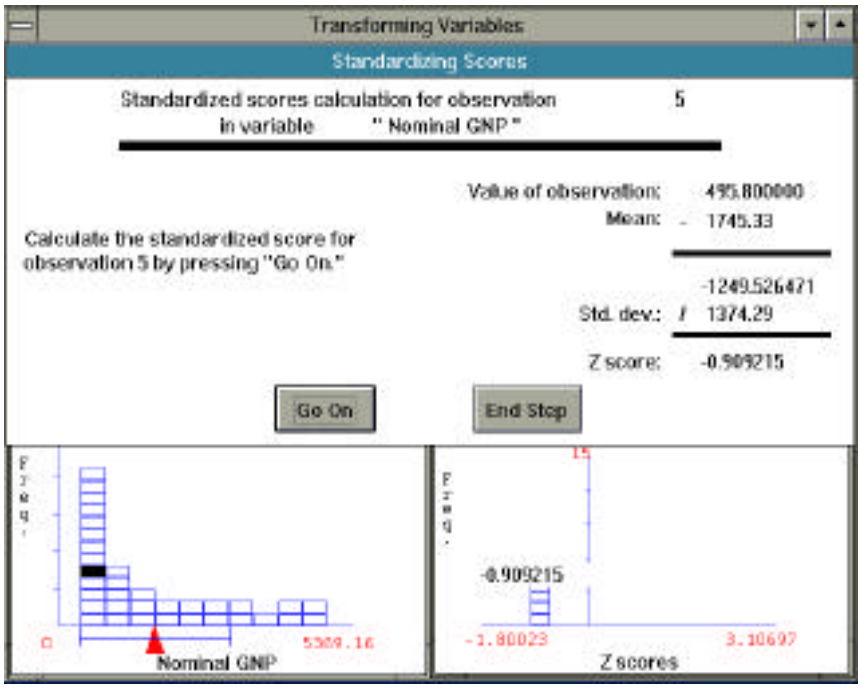

Figure 4: Illustrating a *Z*-score transformation using step-by-step animation

This screen is intended to teach students about *Z* scores. However, it also uses histograms in an instructive way--the animation turns boxes into numbers and replots them into intervals. Students who only partially understood histograms might benefit from this illustration in unintended ways. In addition, the means and

standard deviation, which are more central to *Z* scores than histograms, are displayed graphically in the histogram on the lower left. Finally, this is more than just an experiment; that is, the program is illustrating a process.

Figure 5 shows the next screen with the two histograms side by side after the transformation is complete. Data points highlighted in the left histogram appear in the same location (relative to other data points) in the histogram to the right. The hope is that students studying and interacting with the histograms will see that the *Z*-score transformation has not changed the shape of the distribution. However, it too may help students to learn about histograms.

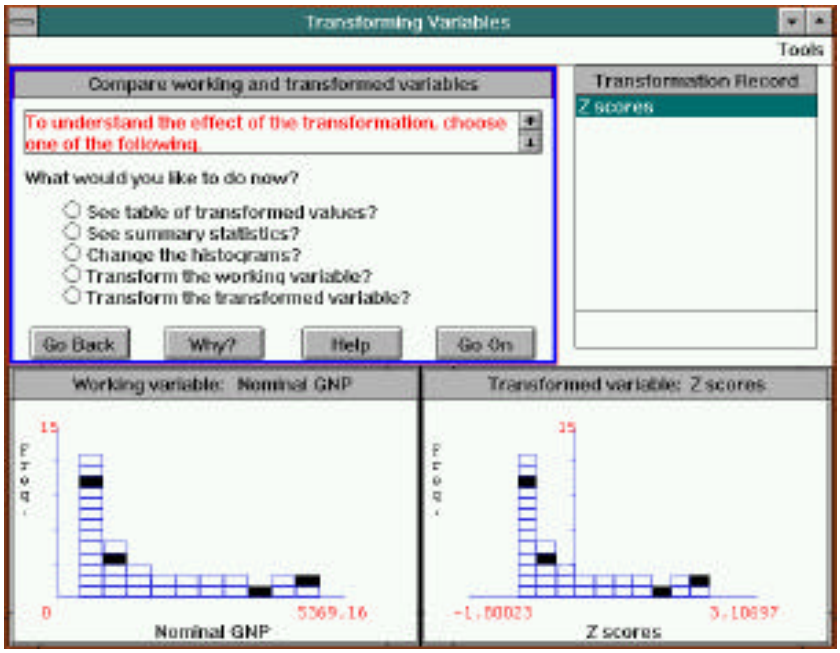

Figure 5: Experimenting after the *Z*-score illustration

Finally, a transformations facility (with *Z* scores and linear and nonlinear transformations) exists as an option in the main pathway of the Describing Bivariate Data program, as well as in the "data analysis" facility shown in Figure 3. In both these pathways, students are using transformations in data analysis more than performing experiments with the goal of learning. Students who follow-up the exercises in Figures 4 and 5 by using transformations in a bivariate analysis might show improved comprehension of *Z* scores and other transformations.

For the development team, isolating concepts meant going through each part of each program and recording each of the comprehension points that the screens might help students learn. There were over 1,000 total comprehension points spread out over the programs, with many redundancies or near redundancies. To make the assessment manageable, we combined redundant and near redundant points into clusters. There were 103 clusters, each tied to specific parts of specific programs. For each cluster, we designed a question to test conceptual understanding.

Questions

For each cluster, we constructed a question that tested conceptual understanding (Cohen, Chechile, Smith, Tsai, & Burns, 1994). All 103 questions were subject to the following criteria: (1) the statistical concept was included in the software, (2) the question was appropriate for an introductory course in statistics, and (3) the question assessed conceptual understanding. Most questions required either near transfer (very similar to the computer exercise) or far transfer (clearly different from the exercise) of conceptual knowledge (Campione & Brown, 1990). The questions were reviewed against these criteria first by internal members of the evaluation team, and then by two outside statistical consultants and professors of quantitative methods. The 103 questions were divided into three tests that covered similar but not identical content. Each test had approximately six questions on the following topics: displaying data, descriptive statistics, transformations, bivariate representation and regression, probability distributions, and sampling. Figure 6 shows one of the questions used to test understanding of *Z* scores.

A university testing center had an established policy of converting raw test scores into standard scores where the mean = 500 and the standard deviation = 100. The computing center of the university recently suggested that the testing center change the standard score system to one with a mean = 1 and standard deviation = 2. What would a score of 420 in the old system be in the new system?

Figure 6: A sample question

Tracing use

To assess individual student use, we added a trace facility in ConStatS (Cohen, Tsai, & Chechile, 1995). The facility permitted us to carryout the assessment without standardizing use and time spent with the programs. In addition to capturing the total time on each program (and screen), the trace facility recorded each keystroke in terms of its purpose. For instance, each time a student clicked on a WHY or HELP button, the interaction was recorded as *Information Retrieval*. When students changed the number of intervals in a histogram, it was recorded as *Experiment*. Experiments were recorded along with relevant parameters (i.e., the number of intervals entered by the student). Finally, *Z*-score transformations, such as the the one illustrated in Figure 4, were recorded as *Animation*. Every keystroke was assigned to a category. The set of categories is described in Cohen et al. (1995).

Summary of participants

As described in Cohen, Smith, Chechile, Burns, and Tsai (1996), 20 different introductory statistics and research methods courses, with 739 students, participated in the assessment over two years. Most of the students were undergraduates. About 62% of the students were women. The courses were taught in seven separate disciplines: psychology, economics, child study, biology, sociology, engineering, and education. Sixteen of the classes (621 students) were taught at the authors' home institution, and four courses (118

students) were at outside colleges and universities. For students using the software, test scores counted for at least 5% of their overall grade in the course. Many instructors added written assignments based on the software that also counted toward the grade. Four classes (77 students) from our home institution participated as control subjects. Each control subject received $50 to participate.

Results of the assessment

Many students showed problems with the basic mathematics skills assumed by the software. In particular, students had problems with two questions: converting .375 to a fraction (missed by 19% of the students), and specifying a ratio between 5:2 and 20:6 (missed by 34%).

Table 1 shows the percent correct on the comprehension test by the number correct on the basic skills test, where 10 is all correct. All students using ConStatS outperformed those in the control classes, and those with basic math skills showed the largest gain. The results showed a similar trend for students at Tufts and for students at the outside institutions.

Table 1: Percent correct on the comprehension test by number correct on the basic skills test

	Number correct on the basic skills test		
	8 or less	9	10
Control	37	41	44
Experimental	46	51	57

Currently, we have not comprehensively interpreted learning outcomes in terms of the trace data. A very preliminary analysis of the trace data showed two questions where specific experimental behavior correlated with higher scores on comprehension test questions (Cohen et al., 1995). For instance, students using ConStatS can experiment with discrete probability distributions by specifying a range of values on the horizontal axis and then learning the probability of observing a value in that range. To assess the effectiveness of this exercise, one question assessed students ability to interpret a discrete probability distribution. Those students who performed experiments with discrete distributions that yielded consecutive non-zero, zero, non-zero probabilities scored much higher on the question than those who did not perform this set of experiments.

Two other interesting outcomes are worth noting, both regarding how the software was integrated into various curriculums:

- One class integrated the software by dropping one class per week and adding a computer lab. They performed as well or better than most classes in the experimental group.
- Nearly every class included some kind of hand-in assignment. Most assignments included specific exercises, questions, and required essays on experiments.

In addition to quantitative analyses of the learning outcomes, a qualitative analysis of student answers yielded 10 patterns of errors on 24 of the questions on the comprehension tests (several patterns appeared

on more than one question) (Cohen et al., 1996). Many of the questions involved interpretation of graphs, particularly histograms, scatterplots, and both cumulative and density plots of probability distributions. For instance, many students providing incorrect (or partially incorrect) answers on questions designed to assess comprehension of histograms seemed confused about the meaning of the vertical axis. They sometimes offered an interpretation more consistent with the vertical axis on a scatterplot (i.e., as representing the values of a typical dependent variable rather than the number of data points falling in a class interval). Similarly, many students offered interpretations of probability distributions that indicated confusion about the difference between probability and data. For instance, some students, when interpreting a normal probability distribution describing the weight of newborn cats, claimed the distribution "did not account for outliers." Of course, the normal probability distribution extends to infinity and "does account for outliers." The students' incorrect answers seem more consistent with an interpretation of a finite distribution of data with distinct low and high observed values.

Thus, while students in all classes using ConStatS showed improvement over those students in the control classes, remedial problems with basic mathematics and confusing properties of displays limited improvement. Even those students with adequate basic mathematical skills still scored only an average of 57% on the test of conceptual understanding. The move to working with data and experimental learning with instructional software can benefit statistics education, but the transition needs to be undertaken with caution. As technology becomes a more central part of teaching the conceptual side of statistics, issues about the use of graphs and optimal experiments will need to be addressed. Having students work with displays rather than symbolic notations offers both advantages and new problems. The problems are best discovered by detailed assessments.

REFERENCES

Campione, J., & Brown, A. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Fredericksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, NJ: Erlbaum.

Clark, R. (1994). Assessment of distance learning technology. In E. Baker & H O'Neill, Jr. (Eds.), *Technology assessment in education and training* (pp. 63-78). Hillsdale, NJ: Erlbaum.

The Cognition and Technology Group at Vanderbilt. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher, 19*(6), 2-10.

Cohen, S., Chechile, R., Smith, G., Tsai, F., & Burns, G. (1994). A method for evaluating the effectiveness of educational software. *Behavior Research Methods, Instruments & Computers, 26* (2), 236-241.

Cohen, S., Smith, G., Chechile, R., Burns, G., & Tsai, F. (1996). Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics, 21*(1), 35-54.

Cohen, S., Smith, G., Chechile, R., & Cook, R. (1994). Designing software for conceptualizing statistics. *Proceedings of the First Conference of the International Association for Statistics Education*.

Cohen, S., Tsai, F., & Chechile, R. (1995). A model for assessing student interaction with educational software. *Behavior Research Methods, Instruments & Computers, 27*(2), 251-256.