ASSESSING STUDENTS' CONCEPTUAL UNDERSTANDING AFTER A FIRST COURSE IN STATISTICS

ROBERT DELMAS
University of Minnesota
delma001@umn.edu

JOAN GARFIELD
University of Minnesota
jbg@umn.edu

ANN OOMS
University of Minnesota
ooms0001@umn.edu

BETH CHANCE
California Polytechnic State University
bchance@calpoly.edu

Paper presented at the Annual Meetings of
The American Educational Research Association
San Francisco, CA
April 9, 2006

The Research Presented in this Paper
was Supported by the National Science Foundation
NSF CCLI ASA- 0206571

ABSTRACT

This paper describes the development of the CAOS test, designed to measure students' conceptual understanding of important statistical ideas at the end of an introductory course in statistics. Over a three year period items were developed, revised and tested. Three rounds of evaluation by content experts indicated that the current instrument, a 40 item multiple-choice test, has content validity for students enrolled in a college-level non-mathematical first course in statistics. A reliability analysis found reasonably high internal consistency for the same population of students. Results are reported from a large scale class testing that involved undergraduate students in the first statistics course. Item responses were compared from pretest to posttest for a subset of the students in the class testing in order to learn more about areas in which students demonstrated improved performance from beginning to end of the course, as well as areas that showed no improvement or decreased performance. Items that showed an increase in students' misconceptions about particular statistical concepts were also examined. The paper concludes with a discussion of implications for students' understanding of different statistical topics, followed by suggestions for further research.

INTRODUCTION

What do students know at the end of a first course in statistics? How well do they understand the important concepts, and use basic statistical literacy to read and critique information in the world around them? Students' difficulty with understanding probability and reasoning about chance events is well documented (Garfield, 2003; Konold, 1989, 1995; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Pollatsek, Konold, Well, & Lima, 1984; Shaugnessy, 1977, 1992). Studies indicate that students also have difficulty with reasoning about distributions and graphical representations of distributions (e.g., Bakker & Gravemeijer, 2004; Biehler, 1997; Ben-Zvi 2004; Hammerman & Rubin, 2004; Konold, 2003; McClain, Cobb, & Gravemeijer, 2000), and understanding concepts related to statistical variation such as measures of variability (delMas & Liu, 2005; Mathews & Clark, 1997; Shaugnessy, 1977), sampling variation (Reading & Shaughnessy, 2004; Shaughnessy, Watson, Moritz, & Reading, 1999), and sampling distributions (delMas, Garfield, & Chance, 1999; Rubin, Bruce, & Tenney, 1990; Saldanha & Thompson, 2001). There is evidence that instruction can have positive effects on students' understanding of these concepts (e.g., delMas & Bart, 1989; Lindman & Edwards, 1961; Meletiou-Mavrotheris & Lee, 2002; Sedlmeier, 1999), but many students can still have conceptual difficulties even after the use of innovative instructional approaches and software (Chance, delMas, & Garfield, 2004; Hodgson, 1996; Saldanha & Thompson, 2001).

Partially in response to the difficulties students have with learning and understanding statistics, a reform movement was initiated in the early 1990s to transform the teaching of statistics at the introductory level (e.g., Cobb, 1992; Hogg, 1992). Moore (1997) described the reform movement as primarily having made changes in content, pedagogy, and technology. As a result, Scheaffer (1997) observed that there is more agreement today among statisticians about the content of the introductory course than in the past. Garfield (2001), in a study conducted to evaluate the effect of the reform movement, found that many statistics instructors are aligning their courses with reform recommendations regarding technology, and, to some extent, with teaching methods and assessment. While there is evidence of changes in statistics instruction, a large national study has not been conducted on whether these changes have had a positive effect on students' statistical understanding, especially with difficult concepts like those mentioned above.

One reason for the absence of research on the effect of the statistics reform movement may be the lack of a standard assessment instrument. Such an instrument would need to measure generally agreed upon content and learning outcomes, and be easily administered in a variety of institutional and classroom settings. Many assessment instruments have consisted of teachers' final exams that are often not appropriate if they focus on procedures, definitions, and skills, rather than conceptual understanding (Garfield & Chance, 2000). The Statistical Reasoning Assessment (SRA; Garfield, 2003) was one attempt to develop and validate a measure of statistical reasoning, but it focuses heavily on probability, and lacks items related to data production, data collection, and statistical inference. The Statistics Concepts Inventory (SCI) was developed to assess statistical understanding but it was written for a specific audience of engineering students in statistics (Rhoads, Murphy, & Terry, 2006). The ARTIST project (Garfield, delMas, & Chance, 2002) was funded by NSF to develop an assessment instrument that would have broader coverage of both the statistical content typically covered in the first, non-mathematical statistics course, and apply to the broader range of students who enroll in these courses.

THE ARTIST PROJECT

The National Science Foundation (NSF) funded the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (DUE-0206571) to address the assessment challenge in statistics education as presented by Garfield and Gal (1999), who outlined the need to develop reliable, valid, practical, and accessible assessment items and instruments. The ARTIST Web site (https://app.gen.umn.edu/artist/) now provides a wide variety of assessment resources for evaluating students' statistical literacy (e.g., understanding words and symbols, being able to read and interpret graphs and terms), statistical reasoning (e.g., reasoning with statistical information), and statistical thinking (e.g., asking questions and making decisions involving statistical information). These resources were designed to assist faculty who teach statistics, across various disciplines (e.g., mathematics, statistics, and psychology), in assessing student learning of statistics, to better evaluate individual student achievement, to evaluate and improve their courses, and to assess the impact of reform-based instructional methods on important learning outcomes.

DEVELOPMENT OF THE CAOS TEST

An important component of the ARTIST project was the development of an overall Comprehensive Assessment of Outcomes in Statistics (CAOS). The intent was to develop a set of items that students, completing any introductory statistics course, would be expected to understand. The CAOS test was developed through a three-year process of acquiring and writing items, revisions, feedback from advisors and class testers, and two large content validity assessments. During this process the ARTIST team developed and revised items and the ARTIST advisory board provided valuable feedback as well as validity ratings of items, which were used to determine and improve content validity for the targeted population of students (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

The ARTIST advisory group initially provided feedback and advice on the nature and content of such a test. Discussion led to the decision to focus the instrument on different aspects of reasoning about variability, which was viewed as the primary goal of a first course. This included reasoning about variability in distributions, in comparing groups, in sampling, and in sampling distributions. The ARTIST team had developed an online assessment item database with over 1000 items as part of the project. Multiple choice items to be used in the CAOS test were initially selected from the ARTIST item data base or were created. All items were revised to ensure they involved real or realistic contexts and data, and to ensure that they followed established guidelines for writing multiple choice items (Haladyna, Downing, & Rodriguez, 2002). The first set of items was evaluated by the ARTIST advisory group, who provided ratings of content validity and identified important concepts that were not measured by the test. The ARTIST team revised the test and created new items to address missing content. An online prototype of CAOS was developed during summer 2004, and the advisors engaged in another round of validation and feedback in early August, 2004. This feedback was then used to produce the first version of CAOS, which consisted of 34 multiple-choice items. This version was used in a pilot study with introductory statistics students during fall 2004. Data from the pilot study were used to make additional revisions to CAOS, resulting in a second version of CAOS that consisted of 37 multiple choice items.

The second version, called CAOS 2, was ready to launch as an online test in January, 2005. Administration of the online test required a careful registration of instructors, a means for students to securely access the test online, and provision for instructors to receive timely feedback of test results. In order to access the online tests, an instructor requested an access code, which was then used by students to take the test online. As soon as the students completed the test, either in class or out of class, the instructor could download two reports of students' data. One was a copy of the test, with percentages filled in for each response given by students, and with the correct answers highlighted. The other report was a spreadsheet with the total percentage correct score for each student.

CLASS TESTING OF CAOS 2

The first large scale class testing of the online instruments was conducted during spring 2005. Invitations were sent to teachers of high school Advanced Placement (AP) and college statistics courses through e-mail lists (e.g., AP listserv, Statistical Education Section of the American Statistics Association). In order to gather as much data as possible, a hard copy version of the test with machine readable bubble sheets was also offered. Instructors signed up at the ARTIST Web site to have their students take CAOS 2 as a pretest and /or a posttest, using either the online or bubble sheet format.

Many instructors registered their students to take the ARTIST CAOS test as a pretest at the start of a course and as a posttest toward the end of the course. Although it was originally hoped that all tests would be administered in a controlled classroom setting, many instructors indicated the need for out of class testing. Information gathered from registration forms also indicated that instructors used the CAOS results for a variety of purposes, namely, to assign a grade in the course, for review before a course exam, or to assign extra credit. Nearly 100 secondary-level students and 800 college-level students participated. Results from the analysis of the spring 2005 data were used to make additional changes, which produced a third version of CAOS (CAOS 3).

EVALUATION OF CAOS 3 AND DEVELOPMENT OF CAOS 4

The third version of CAOS was given to a group of 30 statistics instructors who were faculty graders of the Advanced Placement Statistics exam in June, 2005, for another round of validity ratings. Although the ratings indicated that the test was measuring what it was designed to measure, the instructors also made many suggestions for changes. This feedback was used to add and delete items from the test, as well as to make extensive revisions to produce a final version of the test, called CAOS 4, consisting of 40 multiple choice items. CAOS 4 was administered in a second large scale testing during fall 2005. Results from this large scale, national sample of college-level students are reported in the following sections.

In March 2006, a final analysis of the content validity of CAOS4 was conducted. A group of 18 members of the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) were used as expert raters. These individuals are statisticians who are involved in teaching statistics at the college level, and who are considered experts and leaders in the national statistics education community. They were given copies of the CAOS test which had been annotated to show what each item was designed to measure. After reviewing the annotated test, they were asked to respond to a set of questions about the validity of the items and instrument for use as an outcome measure of student learning after a first course in statistics. There was unanimous agreement by the expert raters that CAOS 4 measures important basic learning outcomes, and 94% agreement that it measures important learning outcomes. In addition, all raters agreed with the statement “CAOS measures outcomes for which I would be disappointed if they were not achieved by students who succeed in my statistics courses.” Although some raters indicated topics that they felt were missing from the scale, there was no agreement among these raters about the topics that were missing. Based on this evidence, the assumption was made that CAOS4 is a valid measure of important learning outcomes in a first course in statistics.

CLASS TESTING OF CAOS 4

Description of the Sample

In the fall of 2005, CAOS 4 was administered as an online and hard copy test for a final round of class testing and data gathering for psychometric analyses. A total of 1028 students completed CAOS 4 as a posttest. Several criteria were used to select students from this larger pool as a sample with which to conduct a reliability analysis of internal consistency. To be included in the sample, students had to respond to all 40 items on the test and either have completed CAOS 4 in an in class, controlled setting or, if the test was taken out of class, have taken at least 10 minutes, but no more than 60 minutes, to complete the test. The latter criterion was used to eliminate students who did not engage sufficiently with the test questions or who spent an excessive amount of time on the test, possibly looking up answers.

A total of 817 introductory statistics students, taught by 28 instructors from 25 higher education institutions from 18 states across the United States met these criteria and were included in the sample (see Table 1). The majority of the students whose data were used for the reliability analysis were enrolled at a university or a four-year college, with less than a fifth of the students enrolled in two-year or technical colleges. A little more than half of the students (55%) were females, and 71% of the students were Caucasian.

Table 2 shows the mathematics requirements for entry into the statistics course in which students enrolled. The largest group was represented by students in courses with a high school algebra requirement, followed by a college algebra requirement and no mathematics requirement, respectively. Only 4% of the students were enrolled in a course with a calculus prerequisite.

The majority of the students (66%) took the CAOS posttest in class. Only four instructors used the CAOS test results as an exam score, which accounted for 11% of the students. The most common uses of the CAOS posttest results were to assign extra credit (26%), or for review prior to the final exam (20%), or both (17%).

Table 1

Number of higher education institutions, instructors, and students per institution type for students who completed the CAOS posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	5	5	153	18.7
4 year college	8	10	311	38.1
University	12	13	353	43.2
Total	25	28	817	

Table 2

Number and percent of students per course type

Mathematics prerequisite	Number of students	Percent of students
No mathematics requirement	237	29.0
High school algebra	304	37.2
College algebra	243	29.7
Calculus	33	4.0

Reliability Analysis

Using the sample of students described above, the CAOS test appeared to have acceptable internal consistency for students enrolled in college-level, non-mathematical introductory statistics courses. An analysis of internal consistency of the 40 items on the CAOS posttest produced a Cronbach's alpha coefficient of .77. A second set of analyses looked at possible subscales of the CAOS test, but the alpha coefficients were too low to assume that the test was comprised of separate scales.

Analysis of Pretest to Posttest Changes

A major question that needs to be addressed is whether students enrolled in a first statistics course make significant gains from pretest to posttest on the CAOS test. Total percent correct scores from a subset of students who completed CAOS as both a pretest (at the beginning of the course) and as a posttest (at the end of the course) were compared for 488 introductory statistics students.

Description of the Sample

The 488 students in this sample of matched pretests and posttests were taught by 18 instructors at 16 higher education institutions from 14 states across the United States (see Table 3). University students made up the largest group, followed closely by four-year college students. A little more than 10% of the students were from two-year or technical colleges. The majority of the students were females (61%), and close to three-fourths of the students were Caucasian.

Table 4 shows the distribution of mathematics requirements for entry into the statistics courses in which students enrolled. The largest group was represented by students in courses with a high school algebra requirement, followed by no mathematics requirement, and a college algebra requirement, respectively. Only about 4% of the students were enrolled in a course with a calculus prerequisite.

Table 3

Number of higher education institutions, instructors, and students per institution type for students who completed both a pretest and a posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	2	2	61	12.5
4 year college	6	7	203	41.6
University	8	9	224	45.9
Total	16	18	488	

Table 4

Number and percent of students per type of mathematics prerequisite

Mathematics Prerequisite	Number of students	Percent of students
No mathematics requirement	141	28.9
High school algebra	225	46.1
College algebra	101	20.7
Calculus	21	4.3

About 72% of the students received the CAOS posttest as an in class administration, with the remainder taking the test online outside of regularly scheduled class time. Only four instructors used the CAOS posttest scores solely as an exam grade in the course, which accounted for 18% of the students. The most common use of the CAOS posttest results for students who took both the pretest and posttest was to assign both extra credit and conduct a review before the final exam (25%), with 15% of the students using the CAOS posttest only for review, and another 7% only receiving extra credit. For the remainder of the students (35%), many of the instructors indicated some other use such as program or course evaluation.

Pretest to Posttest Changes in CAOS Test Scores

There was an increase from an average percent correct of 43.3% on the pretest to an average percent correct of 51.2% on the posttest (se = .573; $t(487) = 13.80$, $p < .001$). While statistically significant, this was only a small average increase of 8 percentage points (95% CI = [6.8,9.0] or 2.7 to 3.6 of the 40 items), and it was surprising to find that students were correct on only half the items, on average, by the end of the course. To further investigate what could account for the small gain, student responses on each item were compared to see if there were items with significant gains, items that showed no improvement, or items where the percentage of students with correct answers decreased from pretest to posttest.

Analysis of Pretest to Posttest Changes on Item Responses

The next step in analyzing pretest to posttest gains was to look at changes in correct responses for individual items. Matched-pairs *t*-tests were conducted for each CAOS item to test for statistically significant differences between pretest and posttest percent correct. Responses to each item on the pretest and posttest were coded as 0 for an incorrect response and 1 for a correct response. This produced four different response patterns across the pretest and posttest for each item. An “incorrect” response pattern consisted of an incorrect response on both the pretest and the posttest. A “decrease” response pattern was one where a student selected a correct response on the pretest and an incorrect response on the posttest. An “increase” response pattern occurred when a student selected an incorrect response on the pretest and a correct response on the posttest. A “pre & post” response pattern consisted of a correct response on both the pretest and the posttest. The percent of students who fell into each of these response pattern categories is given in Appendix A.

The change from pretest to posttest in the percent of students who selected the correct response was determined by the difference between the percent of students who fell into the “increase” and “decrease” categories. This becomes a little more apparent if it is recognized that the percent of students who gave a correct response on the pretest was equal to the percent in the “decrease” category plus the percent in the “pre & post” category. Similarly, the percent of students who gave a correct response on the posttest was equal to the percent in the “increase” category added to the percent in the “pre & post” category. When the percent of students in the “decrease” and “increase” categories were about the same, the change tended to not produce a statistically significant effect relative to sampling error. When there was a large difference in the percent of students in these two categories (e.g., one category had twice or more students than the other category), the change had the potential to produce a statistically significant effect relative to sampling error. Comparison of the percent of students in these two “change” categories can be used to interpret the change in percent from pretest to posttest.

A per test Type I Error limit was set at α_c = .001 to keep the study-wide Type I Error rate at α = .05 or less across the 48 paired *t*-tests conducted (see Tables 5 through 9). For each CAOS item that produced a statistically significant change from pretest to posttest, multivariate analyses of variance (MANOVA) were conducted to test for interactions with type of institution and type of mathematics prerequisite. Separate MANOVAs were conducted for these two grouping variables because the two variables were not completely crossed. A *p*-value limit of .001 was again used to control the experiment-wise Type I Error rate. If no interaction was found with either variable, an additional MANOVA was conducted using instructor as a grouping variable, to test if a statistically significant change from pretest to posttest was due primarily to large changes in only a few classrooms.

The following sections describe analyses of items that were grouped into the following categories: (a) those that had high percentages of students with correct answers on both the pretest and the posttest, (b) those that had moderate percentages of correct answers on both pretest and posttest, (c) those that showed the largest increases from pretest to posttest, and (d) those that had low percentages of students with correct responses on both the pretest and the posttest. Tables 5 through 8 present a brief description of what each item assessed, report the percent of students who selected a correct response separately for the pretest and the posttest, and indicate the *p*-value of the respective matched-pairs *t*-statistic for each item.

Items with High Percentages of Students with Correct Responses on both Pretest and Posttest

It was surprising to find several items on which students provided correct answers on the pretest as well as on the posttest. These were eight items on which 60% or more of the students demonstrated an ability or conceptual understanding at the start of the course, and on which 60% or more of the students made correct choices at the end of the course (Table 5). With the exception of item 23, a majority of the students were correct on both the pretest and the posttest for this set of items (see Appendix A). Across the eight items represented in Table 5, about the same percent of students (between 8% and 21%) had a decrease response pattern as had an increase response pattern for each item. The net result was that the change in percent of students who were correct did not meet the criterion for statistical significance for any of these items.

Table 5

Items with 60% or more of students correct on the pretest and the posttest

			Percent of Students Correct		paired t
Item	Measured Learning Outcome	N	Pretest	Posttest	*p*
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.	487	69.0	72.1	.273
11	Ability to compare groups by considering where most of the data are and focusing on distributions as single entities.	483	85.9	84.7	.532
12	Ability to compare groups by comparing differences in averages.	480	84.8	83.5	.616
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	480	59.6	67.7	.002
18	Understanding of the meaning of variability in the context of repeated measurements and in a context where small variability is desired.	472	78.4	74.6	.120
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	470	87.2	88.5	.662
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	473	71.5	77.4	.021
23	Understanding that no statistical significance does not guarantee that there is no effect.	464	61.2	59.9	.657

Around 70% of the students were able to select a correct description and interpretation of a histogram that included a reference to the context of the data (item 1). The most common mistake on the posttest was to select the option that correctly described shape, center, and spread, but did not provide an interpretation of these statistics within the context of the problem.

In general, students demonstrated facility on both the pretest and posttest with using distributional reasoning to make comparisons between two groups (items 11, 12, and 13). Around 85% of the students on the pretest and posttest correctly indicated that comparisons based on single cases were not valid. Students had a little more difficulty with item 13, which required the knowledge that comparing groups does not require equal sample sizes in each group, especially if both sets of data are large. Students appear to have good informal intuitions or understanding of how to compare groups. However, the belief that groups must be of equal size to make valid comparisons is a persistent misunderstanding for some students.

On both the pretest and posttest, students demonstrated a good understanding of variability in the context of repeated measurements and understand that small variability is desired within a specific context (item 18). Item 18 provided a context in which a student was trying to decide on the better of two routes to drive to school. A set of travel times for 5 different days on each route were presented. The fictitious student wanted to choose the route that allowed her to arrive for class on time, but not too early in order to minimize parking fees. Around three-fourths of the students on both tests made a correct choice of the route with a slightly higher mean, but much less variability.

Students also showed some understanding of bivariate relationships on the pretest, which persisted to the posttest (items 20 and 21). The vast majority of students on the pretest and posttest were able to match a scatterplot to a verbal description of a bivariate relationship. Most students on the pretest could also select a correct description of a bivariate relationship shown in a scatterplot when there is an outlier (or influential point), with nearly 80% doing so on the posttest.

A majority of students on the pretest appeared to understand that statistical significance does not mean that there is no effect (item 23). However, making a correct choice on this item was not as persistent as for the items described above; 40% of the students did not demonstrate this understanding on the posttest.

Items that Showed Increases in Percent of Students with Correct Responses from Pretest to Posttest

There were eight items on which there was a statistically significant increase from pretest to posttest, and at least 50% of the students made a correct choice on the posttest (Table 6). For all eight items, less than half of the students were correct on both the pretest and the posttest (see Appendix A). Whereas between 6% and 17% of the students had a decrease response pattern across the items, there were two to four times as many students with an increase response pattern for each item. This resulted in statistically significant increases from pretest to posttest in the percent of students who chose correct responses for each item.

Item 2 asked students to identify a boxplot that represented the same data displayed in a histogram. Performance was around 45% of students correct on the pretest with posttest performance just under 60%. Around half of the students on the pretest were able to match a histogram to a description of a variable expected to have a distribution with a negative skew (item 3), a variable expected to have a symmetric, bell-shaped distribution (item 4), and a variable expected to have a uniform distribution (item 5), with increases of 14 to 16 percentage points from pretest to posttest across the three items. A little better than two-fifths of the students correctly indicated that a small *p*-value is needed to establish statistical significance (item 19), and this increased by 25 percentage points on the posttest. None of the MANOVAs conducted for these five items produced statistically significant interactions with type of institution, type of mathematics prerequisites, or instructor.

Table 6

Items with 50% or more of students correct on the posttest and statistically significant gain

			Percent of Students Correct		paired t
Item	Measured Learning Outcome	N	Pretest	Posttest	*p*
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	485	44.9	57.3	<.001
3	Ability to visualize and match a histogram to a description of a variable (negatively skewed distribution for scores on an easy quiz).	485	51.3	67.2	<.001
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	481	45.3	59.3	<.001
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution for the last digit of phone numbers sampled from a phone book).	483	50.5	64.4	<.001
19	Understanding that low *p*-values are desirable in research studies.	462	43.1	68.2	<.001
29	Ability to detect a misinterpretation of a confidence level (percent of population data values between confidence limits).	462	32.9	62.8	<.001
31	Ability to correctly interpret a confidence interval.	459	49.2	74.7	<.001
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	465	53.8	64.7	<.001

On the pretest, only a third of the students recognized an invalid interpretation of a confidence interval as the percent of the population data values between the confidence limits (item 29), which increased to just under two-thirds on the posttest. There was a statistically significant interaction between the pretest to posttest gain and type of institution ($F(2,459) = 8.38, p < .001$). Post hoc simple effects analysis (Howell, 2002) did not produce a statistically significant effect for type of institution on the pretest ($F(2, 909.26) = 2.43, p = .089$), but did produce a statistically significant effect for type of institution on the posttest ($F(2, 909.26) = 5.59, p = .004$). All three institution types had an increase between 22 and 60 percentage points from pretest to posttest in percent of students who got item 29 correct. The 60 percentage point increase for the technical or two-year colleges was considerably larger than for the other types of institutions, which appears to account for the interaction.

About half of the students recognized a valid interpretation of a confidence interval on the pretest (item 31), which increased to three-fourths on the posttest. Finally, while a little more than half of the students could correctly identify a plausible random sample taken from a population on the pretest, this increased by 11 percentage points on the posttest (item 34). While these students showed both practical and statistically significant gains on all of the items in Table 6, anywhere from 26% to 40% still did not make the correct choice for this set of items on the posttest. None of the MANOVAs conducted for these two items produced statistically significant interactions.

There were six additional items that produced statistically significant increases in percent correct from pretest to posttest, but where the percent of students with correct responses on the posttest was still below 50% (Table 7). Similar to the items in Table 6, between 8% and 17% of the students had a decrease response pattern. However, for each item, about twice as many students had a response pattern that qualified as an increase. The net result was a statistically significant increase in the percent of students correct for all six items.

Table 7

Items with less than 50% of students correct on the posttest and statistically significant gain

			Percent of Students Correct		paired t
Item	Measured Learning Outcome	N	Pretest	Posttest	*p*
6	Understanding to properly describe the distribution of a quantitative variable (shape, center, and spread), need a graph like a histogram which places the variable along the horizontal axis and frequency along the vertical axis.	482	13.1	22.6	<.001
10	Understanding of the interpretation of a median in the context of boxplots.	483	18.0	25.7	.001
14	Ability to correctly estimate and compare standard deviations for different histograms. Understands lowest standard deviation would be for a graph with the least spread (typically) away from the center.	476	31.5	46.4	<.001
16	Understanding that statistics from small samples vary more than statistics from large samples.	474	21.7	29.5	.001
38	Understanding of the factors that allow a sample of data to be generalized to the population.	452	21.9	37.2	<.001
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	456	37.3	47.6	.001

In general, students demonstrated some difficulty interpreting graphic representations of data. On item 6, less than one-fourth of the students on the pretest and the posttest demonstrated the understanding that a graph like a histogram is needed to show shape, center and spread of a distribution of quantitative data. The nearly 10 percentage point increase from pretest to posttest in percent of students selecting the correct response was statistically significant. Most students (45% on the pretest and 55% on the posttest) selected a bar graph with a bell-shape, but such a graph could not be used to directly determine the mean, variability, and shape of the measured variable. Students demonstrated a tendency to select an apparent bell-shaped or normal distribution, even when this did not make sense within the context of the problem.

The MANOVAs conducted for item 6 responses with type of institution and type of mathematics preparation did not produce significant interactions. The MANOVA that included instructor as an independent variable did produce a statistically significant interaction between pretest to posttest change and instructor ($F(17, 464) = 2.58$, $p = .001$). A post hoc simple effects analysis (Howell, 2002) did not produce a statistically significant effect for instructor on the pretest ($F(17, 861.29) = .89$, $p = .587$), but did produce a statistically significant effect for instructor on the posttest ($F(17, 861.29) = 4.62$, $p < .001$). The majority of the courses (12 of the 18 instructors) showed increases of 4 to 50 percentage points on item 6 from pretest to posttest, with the students for four instructors producing decreases (4 to 15 percentage points), and the students for two instructors remaining essentially the same. The interaction may be partially due to the students of four instructors producing relatively larger gains of 27 to 50 percentage points, while the other nine courses had gains between 4 and 18 percentage points. Therefore, the overall trend was for an increase in percent correct from pretest to posttest, although the gain was relatively small for most courses when it occurred.

A very small percent of students demonstrated a correct understanding of the median in the context of a boxplot (item 10) on the pretest, with only a small, but statistically significant, improvement on the posttest. Item 10 presented two boxplots positioned one above the other on the same scale. Both boxplots had the same range and median. The width of the box for one graph was almost twice the width of the other graph, with consequently shorter whiskers. On the posttest, most students (66%) chose a response that indicated that the boxplot with a longer upper whisker would have a higher percentage of data above the median. A significant interaction was produced for pretest to posttest change by institution ($F(2, 480) = 7.26$, $p = .001$). The average percent correct increased for four-year institutions or universities (16 and 5 percentage point increases, respectively), whereas there was a decrease from pretest to posttest for the technical or two-year college students (11 percentage point decrease). In 16 of the 18 courses, less than 45% of the students gave a correct response on the posttest.

Item 14 asked students to determine which of several histograms had the lower standard deviation. Just under half of the students answered this item correctly on the posttest. The 15 percentage point increase in percent correct from pretest to posttest, however, was statistically significant.

Item 16, required the understanding that statistics from relatively small samples vary more than statistics from larger samples. Although the increase was statistically significant ($p < .001$), only about one-fifth of the students answered this item correctly on the pretest and less than a third did so on the posttest. None of the MANOVA analyses for item 16 produced statistically significant interactions. A slight majority of students (59% on the pretest and 51% on the posttest) indicated that both sample sizes had the same likelihood of producing an extreme value for the statistic.

Many students did not demonstrate a good understanding of sampling principles. Only one-fifth of the students on the pretest, and nearly 40% on the posttest made a correct choice of conditions that allow generalization from a sample to a population (item 38). Even though this was a statistically significant gain from pretest to posttest, over 56% indicated that a random sample of 500 students presented a problem for generalization (supposedly because it was too small a sample to represent the 5000 students living on campus). No statistically significant interactions were produced by the MANOVA analyses.

Less than half of the students could identify a correct interpretation of rejecting the null hypothesis (item 40) on the posttest. While there was a statistically significant gain in correct responses from pretest to posttest, a little over a third of the students indicated that rejecting the null hypothesis meant that it was definitely false, which was 10 percentage points higher than the percent who gave this response on the pretest.

Items with Low Percentages of Students with Correct Responses on Both the Pretest and the Posttest

Table 8 shows that for about a third of the items on the CAOS test less than 50% of the students were correct on the posttest with the change from pretest to posttest not statistically significant, despite having experienced the curriculum of a college-level first course in statistics. Across all of these items, similar percents of students (between 7% and 25%) had a decrease response pattern as had an "increase" response pattern (see Appendix A). The overall result was that none of the changes from pretest to posttest in percent of students selecting a correct response were statistically significant.

Students had very low performance, both pretest and posttest, on item 7 which required an understanding for the purpose of randomization (to produce treatment groups with similar characteristics). On the posttest, about a third of the students chose "to increase the accuracy of the research results" and another third chose "to reduce the amount of sampling error."

Table 8

Items with less than 50% of students correct on the posttest and gain not statistically significant

			Percent of Students Correct		paired t
Item	Measured Learning Outcome	N	Pretest	Posttest	*p*
7	Understanding of the purpose of randomization in an experiment.	483	9.7	13.5	.049
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles.	481	22.2	22.2	1.000
15	Ability to correctly estimate standard deviations for different histograms. Understands highest standard deviation would be for a graph with the most spread (typically) away from the center.	476	41.8	46.4	.116
17	Understanding of expected patterns in sampling variability.	476	40.1	45.8	.031
22	Understanding that correlation does not imply causation.	470	51.9	49.4	.377
28	Ability to detect a misinterpretation of a confidence level (the percent of sample data between confidence limits)	463	48.4	40.6	.009
30	Ability to detect a misinterpretation of a confidence level (percent of all possible sample means between confidence limits)	460	35.0	42.2	.023
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	459	15.7	19.4	.119
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	464	42.5	36.2	.046
35	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.	454	36.1	43.6	.013
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	456	49.6	47.4	.478
37	Understanding of how to simulate data to find the probability of an observed value.	461	22.1	19.7	.314
39	Understanding of when it is not wise to extrapolate using a regression model.	447	22.1	32.4	.930

Students demonstrated some difficulty with understanding how to correctly interpret boxplots. Item 9 was based on the same two boxplots presented for item 10 (Table 7). Only a fifth of the students demonstrated an understanding that boxplots do not provide estimates for percentages of data above or below values except for the quartiles. The item asked students to indicate which of the two boxplots had a greater percentage of cases at or below a specified value. The value did not match any of the quartiles or extremes marked in either boxplot, so the correct response was that it was impossible to determine. The correct response rate was close to chance level on both the pretest and posttest. Sixty percent of students on the posttest indicated that the boxplot with the longer lower whisker had a higher percentage of cases below the indicated value, similar to the erroneous response to item 10. On the posttest, 49% of the students selected the identified erroneous responses to both items 9 and 10.

Item 15 asked students to determine which of several histograms had the highest standard deviation. Similar to item 14 (Table 7), a little under half of the students answered this item correctly on the posttest. There was about a five percent increase in percent correct from pretest to posttest, but the difference was not statistically significant.

Item 17 presented possible results for 5 samples of equal sample size taken from the same population. Less than half the students on the pretest and posttest chose the sequence that represented the expected sampling variability in the sample statistic. A noticeable percent of students on the pretest (38%) and the posttest (35%) indicated that all three sequences of sample statistics were just as plausible, even though one sequence showed an extreme amount of sampling variability given the sample size, and another sequence presented the same sample statistic for each sample (i.e., no sampling variability). In addition, 60% of the students who gave an erroneous response to item 17 on the posttest also selected an erroneous response for item 16

(Table 7), which was discussed earlier as another item where students demonstrated difficulty with understanding sampling variability.

Two other items related to sampling variability proved difficult for students. Item 32 required students to recognize that an estimate of sampling error was needed to conduct an informal inference about a sample mean. Less than 20% of the students made a correct choice on the pretest and posttest. A slight majority of the students (54% pretest, 55% posttest) chose the option which based the inference solely on the sample standard deviation, not taking sample size and sampling variability into account. Item 35 asked students to select a graph from among three histograms that represented a sampling distribution of sample means for a given sample size. Slightly more than a third did so correctly on the pretest, with a slight increase on the posttest.

While it was noted earlier that students could correctly identify a scatterplot given a description of a relationship between two variables, they did not perform as well on another item related to interpreting correlation. About half of the students chose a response indicating that a statistically significant correlation establishes a causal relationship (item 22).

Students did not demonstrate a firm grasp of how to interpret confidence intervals. There was an increase in the percent of students who incorrectly indicated that the confidence level represents the expected percent of sample values between the confidence limits (item 28), although the difference was not statistically significant. Similarly, the majority of students on the posttest indicated that the confidence level indicated the percent of all sample means that fall between the confidence limits (item 30).

Item 33 required the understanding that a distribution with a median greater than the mean is most likely skewed to the left. There was a decrease, though not statistically significant, in the number of students who demonstrated this understanding. The percent of those who incorrectly selected a somewhat symmetric, mound-shaped bar graph increased from 53% on the pretest to 60% on the posttest. A slight majority (59%) of those who made this choice on the posttest also incorrectly chose the bell-shaped bar graph for item 6 (Table 7) discussed earlier.

Slightly less than half of the students correctly indicated that ratios based on marginal totals were needed to make comparisons between rows in a two-way table of counts (item 36), with a little over one-third selecting proportions based on the overall total count on the posttest. Only a fifth of the students indicated that it is not appropriate to extrapolate a regression model to values of the predictor variable that are well beyond the range of values investigated in a study (item 39).

Eighty percent of the students did not demonstrate knowledge of how to simulate data to estimate the probability of obtaining a value as or more extreme than an observed value (item 37). In a situation where a person has to predict between two possible outcomes, the item asked for a way to determine the probability of making at least four out of six correct predictions just by chance. On the posttest, forty-five percent of the students indicated that repeating the experiment a large number of times with a single individual, or repeating the experiment with a large group of people and determining the percent who make four out of six correct predictions, were equally effective as calculating the percent of sequences of six trials with four or more correct predictions for a computer simulation with a 50% chance of a correct prediction on each trial.

Item Responses that Indicated Increased Misconceptions and Misunderstandings

While the items discussed in the previous section showed a drop in the percent of students with correct responses from pretest to posttest, none of these differences were statistically significant. There were, however, several items that showed a statistically significant increase from pretest to posttest in the percent of students selecting a specific erroneous response (Table 9). None of these responses produced statistically significant interactions between pretest to posttest increases and either type of institution, type of mathematics preparation, or instructor. Most of these misunderstandings and misconceptions were discussed in earlier presentations of the results. They include selecting a bell-shaped bar graph to represent the distribution of a quantitative variable (item 6), confusing random assignment with random sampling (item 7), selecting a histogram with a larger number of different values as having a larger standard deviation (item 15), inferring causation from correlation (item 22), use of grand totals to calculate conditional probabilities (item 36), and indicating that rejecting the null hypothesis means the null hypothesis is definitely false (item 40).

Across this set of items, 13% to 17% of the students had a decrease response pattern with respect to the identified erroneous response (see Appendix B). For each item, between one and a half to two times as many students had an increase response pattern with respect to giving the erroneous response. The result was a statistically significant increase in the percent of students selecting the identified responses for each item. Together, these increases indicate that a noticeable number of students developed misunderstandings or misconceptions by the end of the course that they did not demonstrate at the beginning.

Table 9

Items with an increase in a misconception or misunderstanding from pretest to posttest.

			Percent of Students		Paired t
Item	Misconception or Misunderstanding	N	Pretest	Posttest	*p*
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	482	45.4	55.4	.001
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	483	36.4	49.7	<.001
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	476	21.4	30.7	.001
22	Causation can be inferred from correlation.	470	26.8	36.4	.001
36	Grand totals are used to calculate conditional probabilities.	456	27.2	37.3	.001
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	456	25.7	35.1	.001

DISCUSSION

What do students know at the end of their first statistics course? What do they gain in reasoning about statistics from the beginning of the course to the end? Those were the questions that guided an analysis of the data gathered during the Fall 2005 Class testing of the CAOS 4 test. It was disappointing to see such a small overall increase in correct responses from pretest to posttest, especially when the test was designed (and validated) to measure the most important learning outcomes for students in a non-mathematical, first course in statistics. It was also surprising that for almost all items, there was a noticeable number of students who selected the correct response on the pretest, but chose an incorrect response on the posttest.

The following three broad groups of items emerged from the analyses: (a) items on which students seemed to do well on both prior to and after their first course, (b) items where they showed the most gains in learning, and (c) items that were more difficult for students to learn. While less than half of the students were correct on the posttest for all items in the latter category, there was a significant increase from pretest to posttest for about one third of the items in this group. Finally, items were examined that showed an increase in misconceptions about particular concepts. The following sections present a discussion of these results, organized by topic areas: data collection and design, descriptive statistics, graphical representations, boxplots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, and tests of significance.

Data Collection and Design

Students did not show significant gains in understanding some important principles of design, namely the purpose of random assignment and that a correlation from an observational study does not allow causal inferences to be drawn. In fact, students' misconceptions increased in terms of believing that random assignment is equivalent to random sampling or that random assignment reduces sampling error.

Descriptive Statistics

Students seemed to initially understand the idea of variability of repeated measures. While a small percent of students made gains in estimating and identifying a histogram with the lowest standard deviation, there were no significant gains in estimating and identifying a graph with the highest standard deviation among a set of histograms. It seems that some students understood that a graph that is very narrow and clumped in the middle might have less variability, but had different ideas about what more variability might look like (e.g., bumpiness rather than spread from the center). One misconception that increased from pretest to posttest was that a graph with the largest number of different values has the larger standard deviation (spread not considered).

Graphical Representations

Most students seemed to recognize a correct and complete interpretation of a histogram when entering their course, and this did not change after instruction. They did make significant gains in being able to match a histogram to a description of a variable. Only a small percentage of students made gains in understanding that shape, center and spread were represented by a histogram and not a bar graph. One of the most difficult items that showed no significant improvement indicated that students failed to recognize that a distribution with a median larger than the mean is most likely skewed left. Most students were able make reasonable comparisons of groups using dot plots, and students appeared to gain in their understanding that equal sample sizes are not needed to compare groups

Boxplots

Students seemed to have many difficulties understanding and interpreting boxplots. A small percent of students made significant gains in recognizing and interpreting the median in the context of a boxplot. On the posttest, many students seemed to think that the boxplot with the longer lower whisker had a higher percentage of cases below an indicated value or that the boxplot with a longer upper whisker would have a higher percentage of data above the median.

There was no apparent gain in students' understanding that boxplots only provide estimates of percentages between quartiles.

Normal Distribution

Students tended to select responses across various items that showed a normal distribution, suggesting a tendency to select a graph that is like a normal distribution regardless of whether it makes sense to do so within the context of the problem. Presented with an item that reported a median that is noticeably greater than the mean, most students selected a more symmetric, bell-shaped histogram instead of a histogram that is skewed to the left. Many students incorrectly selected a somewhat symmetric, mound-shaped bar graph as a graph that would indicate shape, center and spread, rather than a histogram that was not bell shaped.

Bivariate Data

Students seemed to do a good job at the beginning of their courses with matching a scatterplot to a verbal description, indicating that they understood how a positive linear relationship was represented on a scatterplot. However, no significant gains were made in recognizing that it is not legitimate to extrapolate using values outside the domain of values for the independent variable when using a regression model. Of course, it cannot be determined whether the difficulty comes from students not understanding this idea, students not identifying this idea as the focus on the question asked, or the topic not being covered in the course.

Probability

The probability topics presented in the CAOS 4 test were quite difficult for students. Students showed no gains from pretest to posttest on items that required identification of correct ratios to use when constructing probabilities from a two-way table, or knowing how to simulate data to find the probability of an outcome.

Sampling Variability

Students demonstrated difficulty with understanding sampling variability and sampling distributions. There was only a small increase in the percent of students who demonstrated an understanding that statistics from relatively small samples vary more than statistics from larger samples, or an understanding of factors that allow generalization from a sample to a population. Students had the most difficulty (showed no gains from pretest to posttest) on items that had them select a histogram representing a sampling distribution from a given population for a particular sample size, or asked them to use sampling error as an appropriate measure when making an informal inference about a sample mean.

Confidence Intervals

Students did not demonstrate an understanding of confidence intervals. There was an increase in the percent of students who incorrectly indicated that a confidence level represents the expected percent of sample values between the confidence limits. The majority of students on the posttest also incorrectly indicated that a confidence level indicated the percent of all sample means that fall between the confidence limits. While three-fourths of the students recognized a valid interpretation of a confidence interval on the posttest, many of these same students indicated that the invalid statement also applied, as if the two statements had the same interpretation.

Tests of Significance

Many students entered the course already recognizing that lack of statistical significance does not mean no effect. Most students seem to improve from pretest to posttest in understanding that a low *p*-value is required for statistical significance. A small percent of students made gains in identifying a correct interpretation of a hypothesis test when the null hypothesis is rejected. One misconception that increased from pretest to posttest was that rejecting the null hypothesis means that the null hypothesis is definitely false.

SUMMARY

The CAOS test provides valuable information on what students appear to learn and understand on completing a college-level, non-mathematical first course in statistics. Across college-level first courses in statistics at a variety of institutions, there were some concepts and abilities that many students demonstrated at the start of a course. These included recognizing a complete description of a distribution and understanding how bivariate relationships are represented in scatterplots. Most students also demonstrated an ability to make reasonable interpretations of some graphic representations by the end of a course. However, the results indicate that many students do not demonstrate a good understanding of much of the content covered by the CAOS 4 test, content that statistics faculty agreed represents important learning outcomes for an introductory statistics course. At the end of their respective courses, students still had difficulty with identifying appropriate types of graphic representations, especially with interpreting boxplots. They also did not demonstrate a good understanding of important design principles, or of important concepts related to probability, sampling variability, and inferential statistics.

It should be noted that all items on the CAOS test were written to require students to think and reason, not to compute, use formulas, or recall definitions, contrary to many instructor designed exams on which there may be more pretest to posttest gains. However, the CAOS test purposefully designed to be different from the traditional test written by course instructors. During interviews and on surveys conducted to evaluate the ARTIST project, many instructors communicated that they were quite surprised when they saw their students' scores. They reported that they found the CAOS test results quite illuminating, causing them to reflect on their own teaching in light of the test results. That is one of the most important purposes of the CAOS test, to provide information to statistics instructors to allow them to see if their students are learning to think and reason about statistics, and to promote changes in teaching to better promote these learning goals.

The CAOS test is now available for research and evaluation studies in statistics education. Plans are currently underway for the development of a collaborative effort among many institutions to gather online large amounts of test data (including CAOS) and instructional data as a way to promote future research on teaching and learning statistics at the college level. In addition, there is a need to conduct studies that explore particular activities and sequences of activities in helping to improve students' statistical reasoning as they take introductory statistics courses. Given the internal reliability of the CAOS test for students in non-mathematical introductory college statistics courses, and that it has been judged to be a valid measure of important learning outcomes for students enrolled in such courses, we hope that CAOS will facilitate these much needed studies.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer.

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 169-190). Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi & J. B. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 121-146). Dordrecht, Netherlands: Kluwer.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer.

Cobb, G. (1992). Teaching statistics. In *Heeding the Call for Change: Suggestions for Currricular Action*, *MAA Notes, Vol. 22*, 3-33.

Cobb, G. (1993). Reconsidering statistics education: A National Science Foundation conference. *Journal of Statistics Education 1*(1). Retrieved April 2, 2006 at http://www.amstat.org/publications/jse/v1n1/cobb.html.

delMas, R. and Bart, W.M. (1989). The role of an evaluation exercise in the resolution of misconceptions of probability. *Focus on Learning Problems in Mathematics*, *11*(3), 39-54.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3)*,* Retrieved April 2, 2006 at http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55-82. Retrieved April 2, 2005 at http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_delMas_Liu.pdf.

Garfield, J. (2001). *Evaluating the impact of educational reform in statistics: A survey of introductory statistics courses.* Final Report for NSF Grant REC-9732404.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22-38. Retrieved March 16, 2006 at http://www.stat.auckland.ac.nz/~iase/serj/SERJ2 (1).pdf.

Garfield, J. & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning*, *2*, 99-125.

Garfield, J., delMas, R., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project*. Accessed March 26, 2006 at https://app.gen.umn.edu/artist/. NSF CCLI grant ASA- 0206571.

Garfield , J. & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, *67*, 1-12.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41.

Hodgson, T. R. (1996). The effects of hands-on activities on students' understanding of selected statistical concepts. In E. Jakubowski, D. Watkins, & H. Biske (Eds.), *Proceedings of the Eighteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 241–246). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Hogg, R. (1992). Report of workshop on statistics education. In *Heeding the Call for Change: Suggestions for Curricular Action*, *MAA Notes, Vol. 22*, 34-43.

Howell, D. C. (2002). *Statistical Methods for Psychology (Fifth Edition)*. Pacific Grove, CA: Duxbury.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*, 59-98.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, *3*(1). Accessed March 16, 2006 at http://www.amstat.org/publications/jse/v3n1/konold.html.

Konold, C. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., Pollatsek, A., Well, A. D., Lohmeier, J., and Lipson, A. (1993). Inconsistencies in students' Reasoning about probability. *Journal for Research in Mathematics Education, 34*, 392-414.

Lindman, H. and Edwards, W. (1961). Supplementary report: Unlearning the gambler's fallacy. *Journal of Experimental Psychology*, *62*, 630.

Mathews, D., & Clark, J. (1997). Successful students' conceptions of mean, standard deviation, and the Central Limit Theorem. Paper presented at the Midwest Conference on Teaching Statistics, Oshkosh, WI.

McClain, K., Cobb, P., & Gravemeijer, K. (2000). Supporting students' ways of reasoning about data. In M. Burke & F. Curcio (Eds.), *Learning Mathematics for a New Century, 2000 Yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal, 1*(2), 22-37. http://fehps.une.edu.au/serj.

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*, 123-137.

Pollatsek, A., Konold, C., Well, A., and Lima, S. (1984). Beliefs underlying random sampling. *Memory and Cognition, 12*(4), 395-401.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer.

Rhoades, T. R., Murphy, T. J., & Terry, R. (2006). *The Statistics Concept Inventory (SCI)*. Retrieved on March 13, 2006 at http://coecs.ou.edu/sci/.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about Sampling: trouble at the Core of Statistics. Paper presented at the Third International Conference on Teaching Statistics, New Zealand.

Scheaffer, R. (1997). Discussion. *International Statistical Review*, *65*, 156-158.

Sedlmeier, P. (1999). *Improving Statistical Reasoning: Theoretical Models and Practical Implications*. Mahwah, NJ: Erlbaum.

Saldanha, L. A., & Thompson, P. W. (2001). Students' reasoning about sampling distributions and statistical inference. In R. Speiser & C. Maher (Eds.), *Proceedings of The Twenty-Third Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 449-454), Snowbird, Utah. Columbus, Ohio: ERIC Clearinghouse.

Shaugnessy, M. (1977). Misconceptions of probability: From systematic errors to systematic experiments and decisions. In A. Shulte (ed.), *Teaching Statistics and Probability*, *1981 Yearbook* (pp. 90-100), National Council of Teachers of Mathematics.

Shaugnessy, M. (1981). Misconceptions of probability: An experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics*, *8*, 295-316.

Shaugnessy, M. (1992). Research in probability and statistics: Reflections and directions. In A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 465-494).

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidulph & K. Carr (Eds.), *Proceedings of the Twentieth Annual*

Conference of the Mathematics Education Research Group of Australasia (pp. 6-22). Rotorua, N.Z.: University of Waikato.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). School mathematics students' acknowledgment of statistical variation. For the NCTM Research Precession Symposium: *There's More to Life than Centers*. Paper presented at the 77th Annual National Council of Teachers of Mathematics (NCTM) Conference, San Francisco, CA.

APPENDIX A

Percent of students with item response patterns for selected CAOS items

			Item Response Pattern *			
Item	Measured Learning Outcome	N	Incorrect	Decrease	Increase	Pre & Post
1	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.	487	10.3	17.7	20.7	51.3
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	485	25.4	17.3	29.7	27.6
3	Ability to visualize and match a histogram to a description of a variable (negatively skewed distribution for scores on an easy quiz).	485	26.4	6.4	22.3	44.9
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	481	29.3	11.4	25.4	33.9
5	Ability to visualize and match a histogram to a description of a variable (uniform distribution for the last digit of phone numbers sampled from a phone book).	483	27.5	8.1	21.9	42.4
6	Understanding to properly describe the distribution of a quantitative variable (shape, center, and spread), need a graph like a histogram which places the variable along the horizontal axis and frequency along the vertical axis.	482	69.7	7.7	17.2	5.4
7	Understanding of the purpose of randomization in an experiment.	483	89.7	6.8	10.6	2.9
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles.	481	64.2	13.5	13.5	8.7

* Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

APPENDIX A (continued)

			Item Response Pattern *			
Item	Measured Learning Outcome	N	Incorrect	Decrease	Increase	Pre & Post
10	Understanding of the interpretation of a median in the context of boxplots.	483	64.8	9.5	17.2	8.5
11	Ability to compare groups by considering where most of the data are and focusing on distributions as single entities.	483	5.2	10.1	8.9	75.8
12	Ability to compare groups by comparing differences in averages.	480	5.6	10.8	9.8	73.8
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	480	19.8	12.5	20.6	47.1
14	Ability to correctly estimate and compare standard deviations for different histograms. Understands lowest standard deviation would be for a graph with the least spread (typically) away from the center.	476	42.2	11.3	26.3	20.2
15	Ability to correctly estimate standard deviations for different histograms. Understands highest standard deviation would be for a graph with the most spread (typically) away from the center.	476	35.3	18.3	22.9	23.5
16	Understanding that statistics from small samples vary more than statistics from large samples.	474	62.0	8.4	16.2	13.3
17	Understanding of expected patterns in sampling variability.	476	40.5	13.7	19.3	26.5
18	Understanding of the meaning of variability in the context of repeated measurements and in a context where small variability is desired.	472	9.3	16.1	12.3	62.3

* Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

APPENDIX A (continued)

			Item Response Pattern *			
Item	Measured Learning Outcome	N	Incorrect	Decrease	Increase	Pre & Post
19	Understanding that low *p*-values are desirable in research studies.	462	21.0	10.8	35.9	32.3
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	470	3.4	8.1	9.4	79.1
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	473	9.9	12.7	18.6	58.8
22	Understanding that correlation does not imply causation.	470	29.8	20.9	18.3	31.1
23	Understanding that no statistical significance does not guarantee that there is no effect.	464	19.8	20.3	19.0	40.9
28	Ability to detect a misinterpretation of a confidence level (the percent of sample data between confidence limits)	463	35.0	24.4	16.6	24.0
29	Ability to detect a misinterpretation of a confidence level (percent of population data values between confidence limits).	462	26.8	10.4	40.3	22.5
30	Ability to detect a misinterpretation of a confidence level (percent of all possible sample means between confidence limits)	460	38.5	19.3	26.5	15.7
31	Ability to correctly interpret a confidence interval.	459	14.2	11.1	36.6	38.1
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	459	69.5	11.1	14.8	4.6
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	464	37.9	25.9	19.6	16.6

* Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

APPENDIX A (continued)

			Item Response Pattern *			
Item	Measured Learning Outcome	N	Incorrect	Increase	Decrease	Pre & Post
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	465	18.5	16.8	27.7	37.0
35	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.	454	39.4	17.0	24.4	19.2
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	456	29.8	22.8	20.6	26.8
37	Understanding of how to simulate data to find the probability of an observed value.	461	66.2	14.1	11.7	8.0
38	Understanding of the factors that allow a sample of data to be generalized to the population.	452	52.9	10.0	25.2	11.9
39	Understanding of when it is not wise to extrapolate using a regression model.	447	63.3	14.3	14.5	7.8
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	456	35.5	16.9	27.2	20.4

* Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

APPENDIX B

Percent of students with item response patterns for CAOS items assessing misunderstandings and misconceptions

			Item Response Pattern *			
Item	Misconception or Misunderstanding	N	Neither	Decrease	Increase	Pre & Post
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	482	28.5	16.1	26.1	29.3
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	483	35.9	14.4	27.7	22.0
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	476	56.1	13.2	22.5	8.2
22	Causation can be inferred from correlation.	470	50.4	13.2	22.8	13.6
36	Grand totals are used to calculate conditional probabilities.	456	45.6	17.1	27.2	10.1
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	456	49.5	15.4	24.8	10.3

* Neither = did not select the response on either the pretest or the posttest; Decrease = response selected on pretest, but not on the posttest; Increase = response not selected on the pretest, selected on the posttest; Pre & Post = response selected on both the pretest and posttest