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
Kingston University
a.ooms@kingston.ac.uk

BETH CHANCE
California Polytechnic State University
bchance@calpoly.edu

ABSTRACT

This paper describes the development of the CAOS test, designed to measure students' conceptual understanding of important statistical ideas, across three years of revision and testing, content validation, and realiability analysis. Results are reported from a large scale class testing and item responses are compared from pretest to posttest 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.

Keywords: *Statistics education research; Assessment; Conceptual understanding; Online test*

1. 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; Shaughnessy, 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 & Higgins, 2003; McClain, Cobb, & Gravemeijer,

2000), and understanding concepts related to statistical variation such as measures of variability (delMas & Liu, 2005; Mathews & Clark, 1997; Shaughnessy, 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. Although 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) 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 (Garfield, 2003). The Statistics Concepts Inventory (SCI) was developed to assess statistical understanding but it was written for a specific audience of engineering students in statistics (Reed-Rhoads, Murphy, & Terry, 2006). Garfield, delMas, and Chance (2002) aimed 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 would apply to the broader range of students who enroll in these courses.

2. 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.

3. 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 reliable assessment consisting of a set of items that students completing any introductory statistics course would be expected to understand. Given that a reliable assessment could be developed, a second goal was to identify areas where students do and do not make significant gains in their statistical understanding and reasoning.

The CAOS test was developed through a three-year iterative process of acquiring existing items from instructors, writing items for areas not covered by the acquired items, revising items, obtaining feedback from advisors and class testers, and conducting 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 database 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.

3.1. 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 community, 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 2 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).

3.2. EVALUATION OF CAOS 3 AND DEVELOPMENT OF CAOS 4

The third version of CAOS (CAOS 3) 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 CAOS 4 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 4 test that 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 with the statement “CAOS measures basic outcomes in statistical literacy and reasoning that are appropriate for a first course in statistics,” and 94% agreement with the statement “CAOS measures important outcomes that are common to most first courses in statistics.” 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 additional topic

identified by a majority of the raters. Based on this evidence, the assumption was made that CAOS 4 is a valid measure of important learning outcomes in a first course in statistics.

4. CLASS TESTING OF CAOS 4

4.1. DESCRIPTION OF THE SAMPLE

In the fall of 2005 and spring of 2006, CAOS 4 was administered as an online and hard copy test for a final round of class testing and data gathering for psychometric analyses. The purpose of the study was to gather baseline data for psychometric analysis and not to conduct a comparative study (e.g., performance differences between traditional and reform-based curricula). The recruitment approach used for class testing of CAOS 2 was employed, as well as inviting instructors who had given previous versions of CAOS to participate. A total of 1944 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. In addition, students enrolled in high school AP courses were not included in the analysis. Analysis of data from earlier versions of the CAOS test produced significant differences in percentage correct when the AP and college samples were compared. Inclusion of data from AP students might produce results that are not representative of the general undergraduate population, and a comparison of high school AP and college students is beyond the scope of this study.

A total of 1470 introductory statistics students, taught by 35 instructors from 33 higher education institutions from 21 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 about one fourth of the students enrolled in two-year or technical colleges. A little more than half of the students (57%) were females, and 74% of the students were Caucasian.

Table 1. Number of higher education institutions, instructors, and students per institution type for students who completed the CAOS 4 posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	6	6	341	23.1
4-year college	13	14	548	37.3
University	14	15	581	39.5
Total	33	35	1470	

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 3% of the students were enrolled in a course with a calculus prerequisite.

The majority of the students (64%) took the CAOS 4 posttest in class (henceforth refered to as CAOS). Only four instructors used the CAOS test results as an exam score, which accounted for 12% of the students. The most common uses of the CAOS posttest results were to assign extra credit (35%), or for review prior to the final exam (19%), or both (13%).

Table 2. Number and percent of students per course type

Mathematics prerequisite	Number of students	Percent of students
No mathematics requirement	398	27.1
High school algebra	611	41.6
College algebra	420	28.6
Calculus	41	2.8

4.2. RELIABILITY ANALYSIS

Using the sample of students described above, an analysis of internal consistency of the 40 items on the CAOS posttest produced a Cronbach's alpha coefficient of 0.82. Different standards for an acceptable level of reliability have been suggested, with lower limits ranging from 0.5 to 0.7 (see Pedhazur & Schmelkin, 1991). The CAOS test was judged to have acceptable internal consistency for students enrolled in college-level, non-mathematical introductory statistics courses given that the estimated internal consistency reliability is well above the range of suggested lower limits.

5. ANALYIS 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. The total percentage 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 763 introductory statistics students.

5.1. DESCRIPTION OF THE SAMPLE

The 763 students in this sample of matched pretests and posttests were taught by 22 instructors at 20 higher education institutions from 14 states across the United States (see Table 3). Students from four-year colleges made up the largest group, followed closely by university students. Eighteen percent of the students were from two-year or technical colleges. The majority of the students were females (60%), and 77% 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	4	4	138	18.1
4-year college	10	11	395	51.8
University	6	7	230	30.1
Total	20	22	763	

Table 4. Number and percent of students per type of mathematics prerequisite

Mathematics Prerequisite	Number of students	Percent of students
No mathematics requirement	197	25.8
High school algebra	391	51.2
College algebra	161	21.1
Calculus	14	1.8

Sixty-six percent 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 11% of the students. The most common use of the CAOS posttest results for students who took both the pretest and posttest was to assign extra credit (23% of the students). For 22% of the students the CAOS posttest was used only for review, whereas another 16% received extra credit in addition to using CAOS as a review before the final exam. For the remainder of the students (29%), instructors indicated some other use such as program or course evaluation.

5.2. PRETEST TO POSTTEST CHANGES IN CAOS TEST SCORES

There was an increase from an average percentage correct of 44.9% on the pretest to an average percentage correct of 54.0% on the posttest (se = 0.433; $t(762) = 20.98$, $p < 0.001$). Although statistically significant, this was only a small average increase of 9 percentage points (95% CI = [8.2,9.9] or 3.3 to 4.0 of the 40 items). It was surprising to find that students were correct on little more than half of 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.

6. PRETEST TO POSTTEST CHANGES FOR INDIVIDUAL ITEMS

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 percentage 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 percentage of students who fell into each of these response pattern categories is given in Appendix A.

The change from pretest to posttest in the percentage of students who selected the correct response was determined by the difference between the percentage of students who fell into the "increase" and "decrease" categories. This is a little more apparent if it is recognized that the percentage of students who gave a correct response on the pretest was equal to the percentage in the "decrease" category plus the percentage in the "pre & post" category. Similarly, the percentage of students who gave a correct response on the posttest was equal to the percentage in the "increase" category added to the percentage in the "pre & post" category. When the percentage 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 percentage 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 percentage of students in these two "change" categories can be used to interpret the change in percentage from pretest to posttest.

A per test Type I Error limit was set at $\alpha_c = 0.001$ to keep the study-wide Type I Error rate at $\alpha = 0.05$ or less across the 46 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. The dependent variables for each analysis consisted of a 0/1 coded response for a particular item on the pretest and the posttest (0 = incorrect, 1 = correct). The two independent variables for each MANOVA consisted of the pretest/posttest repeated measure and either type of institution or type of mathematics prerequisite. Separate MANOVAs were conducted using only one of the two between-subjects grouping variables because the two variables were not completely crossed. A p-value limit of 0.001 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 see 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 percentage 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.

6.1. 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). A majority of the students were correct on both the pretest and the posttest for this set of items. Across the eight items represented in Table 5, about the same percentage of students (between 5% and 21%) had a decrease response pattern as had an increase response pattern for each item, with the exceptions of items 13 and 21 (see Appendix A). The net result was that the change in percentage 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

Item	Measured Learning Outcome	*n*	% of Students Correct		Paired *t* *p*
			Pretest	Posttest	
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.	760	71.5	73.6	0.266
11	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.	756	88.0	88.2	0.856
12	Ability to compare groups by comparing differences in averages.	753	85.3	85.8	0.741
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	752	61.8	73.5	<0.001
18	Understanding of the meaning of variability in the context of repeated measurements, and in a context where small variability is desired.	746	80.6	80.6	1.00
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	748	90.5	92.5	0.132
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	749	73.6	83.7	<0.001
23	Understanding that no statistical significance does not guarantee that there is no effect.	735	63.1	64.4	0.588

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). Almost 90% 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.

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; a little more than a third of the students did not demonstrate this understanding on the posttest.

6.2. ITEMS THAT SHOWED INCREASES IN PERCENTAGE OF STUDENT WITH CORRECT RESPONSES FROM PRETEST TO POSTTEST

There were seven items on which there was a statistically significant increase from pretest to posttest, and at least 60% of the students made a correct choice on the posttest (Table 6). For all seven items, less than half of the students were correct on both the pretest and the posttest (see Appendix A). Whereas between 6% and 16% of the students had a decrease response pattern across the items, there were two to five times as many students with an increase response pattern for each item, with the exception of item 34. This resulted in statistically significant increases from pretest to posttest in the percentage of students who chose correct responses for each item.

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 about 15 percentage points from pretest to posttest for each of the three items. About half of the students correctly indicated that a small p-value is needed to establish statistical significance (item 19), and this increased by 23 percentage points on the posttest. A significant interaction was produced for pretest to posttest change by instructor ($F(21, 708) = 2.946$, $p < 0.001$). Three instructors had a decrease of seven to 23 percentage points from pretest to posttest, one instructor had essentially no change, 11 instructors had an increase of 10 to 28 percentage points, and seven instructors had an increase of 36 to 63 percentage points. Five of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (19 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

On the pretest, only one third of the students recognized an invalid interpretation of a confidence interval as the percentage of the population data values between the confidence limits (item 29), which increased to around two thirds on the posttest. There

was a statistically significant interaction with instructor [$F(21,703) = 3.163$, $p<.001$]. There was essentially no change in percentage correct from pretest to posttest for three of the instructors. For the other 19 instructors, students showed an increase of 23 to 60 percentage points. The instructor with the highest increase was not the same instructor with the highest increase for item 19. The increase was statistically significant at $p <.001$ for the students of only nine instructors, which could account for the interaction.

Table 6. Items with 60% or more of students correct on the posttest and statistically significant gain

Item	Measured Learning Outcome	*n*	% of Students Correct		Paired *t* *p*
			Pretest	Posttest	
3	Ability to visualize and match a histogram to a description of a variable (negatively skewed distribution for scores on an easy quiz).	760	56.7	73.2	<0.001
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	757	48.0	63.1	<0.001
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).	758	55.9	71.1	<0.001
19	Understanding that low *p*-values are desirable in research studies.	730	49.9	68.5	<0.001
29	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).	725	32.6	67.6	<0.001
31	Ability to correctly interpret a confidence interval.	720	47.1	74.3	<0.001
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	724	55.3	65.2	<0.001

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. There was a statistically significant interaction with instructor [$F(21,698) = 2.787$, $p<.001$]. Students of 20 of the instructors had an increase of 23 to 60 percentage points from pretest to posttest. The students of the other two instructors had a decrease of 7 and 15 percentage changes, respectively, neither of which were statistically significant. The instructor with the highest increase was not the same instructor with the highest increase for either item 19 or item 29. The increase was statistically significant at $p <.001$ for the students of only six instructors, which could account for the interaction.

Finally, athough 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 10 percentage points on the posttest (item 34). Whereas these students showed both practical and statistically significant gains on all of the items in Table 6, anywhere from 26% to 37% still did not make the correct choice for this set of items on the posttest.

There were thirteen additional items that produced statistically significant increases in percentage correct from pretest to posttest, but where the percentage of students with correct responses on the posttest was still below 60% (Table 7). Similar to the items in Table 6, between 7% and 18% of the students had a decrease response pattern. However, for each item, about one and a half to three times as many students had a response pattern that qualified as an increase. The net result was a statistically significant increase in the percentage of students correct for all thirteen items.

In general, students demonstrated some difficulty interpreting graphic representations of data. 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%. 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 10 percentage point increase from pretest to posttest in percentage of students selecting the correct response was statistically significant. Most students (43% on the pretest and 53% on the posttest) selected a bar graph with a bell shape, but such a graph cannot 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 ($F(21, 732) = 3.224$, $p < 0.001$). Only one of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. Two instructors had a small decrease in percentage of students correct from pretest to posttest, three instructors had essentially no change, 12 instructors had an increase of seven to 18 percentage points, and five instructors had an increase of 26 to 47 percentage points. The differential increase in percentage of students who gave a correct response may account for the interaction. Overall, the general trend was for an increase in the percentage of students with correct responses to item 6.

A very small percentage of students demonstrated a correct understanding of the median in the context of a boxplot (item 10) on the pretest, with about a 9% improvement on the posttest. Item 10 presented two boxplots positioned one above the other on the same scale. Both boxplots had the same median and roughly the same range. 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 instructor ($F(21, 732) = 3.958$, $p < 0.001$). Only one of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. Five instructors had a decrease of six to 14 percentage points from pretest to posttest, two instructors had essentially no change, nine instructors had an increase of five to 17 percentage points, and six instructors had an

Table 7. Items with less than 60% of students correct on the posttest, gain statistically significant

Item	Measured Learning Outcome	*n*	% of Students Correct		Paired *t* *p*
			Pretest	Posttest	
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	759	45.5	56.3	<0.001
6	Understanding that to properly describe the distribution (shape, center, and spread) of a quantitative variable, a graph like a histogram is needed.	754	15.1	25.2	<0.001
10	Understanding of the interpretation of a median in the context of boxplots.	754	19.6	28.3	<0.001
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.	746	34.3	51.7	<0.001
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.	747	38.3	46.9	<0.001
16	Understanding that statistics from small samples vary more than statistics from large samples.	747	22.8	31.9	<0.001
17	Understanding of expected patterns in sampling variability.	746	42.8	50.3	<0.001
27	Ability to recognize an incorrect interpretation of a p-value (prob. treatment is effective).	717	42.3	52.7	<0.001
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits).	723	31.4	44.2	<0.001
35	Ability to select an appropriate sampling distribution for a population and sample size.	719	34.5	44.2	<0.001
38	Understanding of the factors that allow a sample of data to be generalized to the population.	715	26.0	37.9	<0.001
39	Understanding of when it is not wise to extrapolate using a regression model.	710	17.9	24.5	0.001
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	716	41.9	52.0	<0.001

increase of 31 to 61 percentage points. Again, the differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (15 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

Item 14 asked students to determine which of several histograms had the lower standard deviation. A little over half of the students answered this item correctly on the posttest. The 17 percentage point increase in percentage correct from pretest to posttest, however, was statistically significant.

Item 15 asked students to determine which of several histograms had the highest standard deviation. Similar to item 14, a little under half of the students answered this item correctly on the posttest. There was about a nine percent increase in percentage correct from pretest to posttest. A significant interaction was found for pretest to posttest change by course type ($F(3, 743) = 5.563$, $p < 0.001$). Simple effects analyses indicated that the change from pretest to posttest was statistically significant increase for students in courses with no mathematics prerequisite ($F(1, 189) = 10.851$, $p = 0.001$) or a high school algebra prerequisite ($F(1, 383) = 16.460$, $p < 0.001$), but not for students in courses with college algebra ($F(1, 158) = 1.872$ $p = 0.173$) or calculus ($F(1, 13) = 1.918$, $p = 0.189$) prerequisites. In fact, the percentage of students correct on item 15 decreased for the latter two groups, although the differences were not 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 ($p < 0.001$), only about one fifth of the students answered this item correctly on the pretest and less than a third did so on the posttest. A slight majority of students on the posttest indicated that both sample sizes had the same likelihood of producing an extreme value for the statistic. A significant interaction was found for pretest to posttest change by type of institution ($F(2, 744) = 7.169$, $p < 0.001$). Simple effects analyses (Howell, 2002) did not produce a significant effect for type of institution on the pretest ($F(2, 1292) = 2.701$, $p = 0.068$), but the effect was significant on the posttest ($F(2, 1292) = 9.639$, $p < 0.001$). Thirty-six percent of students enrolled at a four-year college and 34% of those attending a university gave a correct response on the posttest, whereas only 17% of those enrolled in a technical or two-year college gave a correct response. The percentage of students who gave a correct response was about the same on the pretest and posttest for technical and two-year college students, whereas four-year colleges had a gain of 9 percentage points and universities had a gain of 16 percentage points. Overall, the change in percentage of students who were correct on item 16 was primarily due to students enrolled at four-year institutions and universities.

Item 17 presented possible results for five 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. About one third of students on the pretest (36%) and the posttest (33%) 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, 74% of the students who gave an erroneous response to item 17 on the posttest also selected an erroneous response for item 16.

There was a statistically significant ($p < 0.001$) increase from pretest to posttest in the percentage of students who indicated that the confidence level indicated the percentage of all sample means that fall between the confidence limits (item 30). However, the percentage went from 31% on the pretest to 44% on the posttest, so that the majority of

students did not indicate this understanding by the end of their statistics courses. A significant interaction was produced for pretest to posttest change by instructor ($F(21, 701) = 2.237$, $p < 0.001$). Two instructors had a decrease of 10 and 43 percentage points, respectively, from pretest to posttest, one instructor had essentially no change, 17 instructors had an increase between four and 16 percentage points, and two instructors had an increase of 21 and 37 percentage points, respectively. Three of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced a statistically significant difference at $p < 0.001$. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and others with relatively large increases. Overall, the majority of instructors (19 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

Item 27 presented a common misinterpretation of a p-value as the probability that a treatment is effective. Forty percent of the students answered correctly on the pretest that the statement was invalid, which increased to 53% on posttest. Although the increase was statistically significant, nearly half of the students indicated that the statement was valid at the end of their respective courses.

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 one third did so correctly on the pretest, with 10% more students selecting the correct response on the posttest.

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 62% indicated that a random sample of 500 students presented a problem for generalization on the posttest (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.

Only one fifth of the students indicated on the posttest 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). A significant interaction was produced for pretest to posttest change by instructor ($F(21, 688) = 4.881$, $p < 0.001$). Two of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced statistically significant differences at $p < 0.001$. The two instructors were both from four-year institutions and had increases of 40 and 61 percentage points, respectively. Among the other instructors, four had a decrease of five to 16 percentage points from pretest to posttest, five instructors had essentially no change (between a decrease of five to an increase of five percentage points), seven instructors had an increase of six to 19 percentage points, and four instructors had an increase of 23 to 30 percentage points. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and a few with relatively large increases. Overall, the majority of instructors (13 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

About half of the students could identify a correct interpretation of rejecting the null hypothesis (item 40) on the posttest. Although there was a statistically significant gain in correct responses from pretest to posttest, about one third of the students indicated that rejecting the null hypothesis meant that it was definitely false, which was five percentage points higher than the percentage who gave this response on the pretest. A significant

interaction was produced for pretest to posttest change by instructor ($F(21, 694) = 2.392$, $p < 0.001$). Two of the post hoc simple effects analyses (Howell, 2002) performed for the pretest to posttest change for each instructor produced statistically significant differences at $p < 0.001$. The two instructors were both from four-year institutions and had increases of 39 and 55 percentage points, respectively. Among the other instructors, five had a decrease of five to 22 percentage points from pretest to posttest, five instructors had essentially no change (between a decrease of five to an increase of 4 percentage points), nine instructors had an increase of six to 14 percentage points, and three instructors had an increase of 21 to 39 percentage points. The differential change in the percentage of students who gave a correct response may account for the interaction, with some instructors having a small decrease and a few with relatively large increases. Overall, the majority of instructors (13 out of 22) had an increase from pretest to posttest in the percentage of students with correct responses.

6.3. ITEMS WITH LOW PERCENTAGES OF STUDENTS WITH CORRECT RESPONSES ON BOTH THE PRETEST AND THE POSTTEST

Table 8 shows that for a little less than one third of the items on the CAOS test less than 60% 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 percentages of students (between 6% and 30%) 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 percentage 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 30% of the students chose "to increase the accuracy of the research results," and another 30% chose "to reduce the amount of sampling error."

Students demonstrated some difficulty with understanding how to correctly interpret boxplots. Items 8 and 9 were based on the same two boxplots presented for item 10 (Table 7). Item 8 asked students to identify which boxplot represented a distribution with a larger standard deviation. One boxplot had a slightly larger range (difference of approximately five units) with an interquartile range that was about twice as large as the interquartile range for the other boxplot. Around 59% of the students chose this graph to have a larger standard deviation on the posttest. On item 9, only one 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. Given that item 9 has four response choices, the correct response rate was close to chance level on both the pretest and posttest. Fifty-eight 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, 48% of the students selected the identified erroneous responses to both items 9 and 10.

Table 8. Items with less than 60% of students correct on the posttest, gain not statistically significant

Item	Measured Learning Outcome	*n*	% of Students Correct		Paired *t* *p*
			Pretest	Posttest	
7	Understanding of the purpose of randomization in an experiment.	754	8.5	12.3	0.010
8	Ability to determine which of two boxplots represents a larger standard deviation.	755	54.7	59.2	0.060
9	Understanding that boxplots do not provide accurate estimates for percentages of data above or below values except for the quartiles.	751	23.3	26.6	0.100
22	Understanding that correlation does not imply causation.	743	54.6	52.6	0.371
24	Understanding that an experimental design with random assignment supports causal inference.	731	58.5	59.5	0.689
25	Ability to recognize a correct interpretation of a p-value.	712	46.8	54.5	0.004
26	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).	719	53.1	58.6	0.038
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits).	729	48.4	43.2	0.029
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	718	16.9	17.1	0.883
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	730	41.5	39.7	0.477
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	719	52.7	53.0	0.909
37	Understanding of how to simulate data to find the probability of an observed value.	722	20.4	19.5	0.659

Although 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 one third (36%) of the students chose a response indicating that a statistically significant correlation establishes a causal relationship (item 22). Item 24 required students to understand that causation can be

inferred from a study with an experimental design that uses random assignment to treatments. The percentage of students answering this item correctly on the posttest was just below the threshold of 60%.

Items 25 and 26 measured students' ability to recognize a correct and an incorrect interpretation of a p-value, respectively. There was a noticeable change from pretest to posttest in the percentage of students indicating that item 25 was a valid interpretation, but the difference was just above the threshold for statistical significance. About 55% of the students answered item 5 correctly and 59% answered item 26 correctly on the posttest. Results for these two items, along with item 27, indicate that many students who identified a correct interpretation of a p-value as valid also indicated that an incorrect interpretation was valid. In fact, of the 387 students who answered item 25 correctly on the posttest, only 5% also indicated that the statements for items 26 and 27 were invalid. For the remainder of these students, 56% thought one of the incorrect interpretations was valid, and 39% indicated both incorrect interpretations as valid.

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

An item 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, 59% posttest) chose the option that based the inference solely on the sample standard deviation, not taking sample size and sampling variability into account.

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 percentage of those who incorrectly selected a somewhat symmetric, mound-shaped bar graph increased from 54% on the pretest to 59% on the posttest. Sixty-four percent of those who made this choice on the posttest also incorrectly chose the bell-shaped bar graph for item 6 (Table 7) discussed earlier.

A little more 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). One third of the students incorrectly chose proportions based on the overall total count on the posttest.

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, 46% 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 percentage who make four out of six correct predictions, were equally effective as calculating the percentage 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.

6.4. ITEM RESPONSES THAT INDICATED INCREASED MISCONCEPTIONS AND MISUNDERSTANDINGS

Whereas some of the items discussed in the previous section showed a drop in the percentage of students with correct responses from pretest to posttest, none of these differences was statistically significant. There were, however, several items with noticeable increases from pretest to posttest in the percentage of students selecting a specific erroneous response (Table 9). The change in percentage of students with correct responses was statistically significant for four of the six items in 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).

Table 9. Items with an increase in a misconception or misunderstanding from pretest to posttest

Item	Misconception or Misunderstanding	*n*	% of Students		Paired *t* *p*
			Pretest	Posttest	
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	754	43.0	52.8	<0.001
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	754	36.2	49.2	<0.001
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	747	26.5	33.1	0.002
22	Causation can be inferred from correlation.	743	27.1	35.9	<0.001
36	Grand totals are used to calculate conditional probabilities.	719	25.2	33.4	<0.001
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	716	26.7	32.4	0.015

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 percentage of students selecting the identified responses for four of the items. 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.

7. 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 and Spring 2006 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 that students seemed to do well both prior to and at the end of their first course, (b) items where they showed the most gains in learning, and (c) items that were more difficult for students to learn. Although 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 almost two thirds 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, logically 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.

7.1. 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, the percentage of students demonstrating misconceptions increased in terms of believing that random assignment is equivalent to random sampling, or that random assignment reduces sampling error, or that causation can be inferred from correlation.

7.2. DESCRIPTIVE STATISTICS

Students seemed to initially understand the idea of variability of repeated measures. Whereas a small percentage of students made gains in estimating and identifying the histogram with the lowest standard deviation and the graph with the highest standard deviation among a set of histograms, around half of all the students did not demonstrate this ability on the posttest. 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).

7.3. GRAPHICAL REPRESENTATIONS

Most students seemed to recognize a correct and complete interpretation of a histogram when entering the 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. There was a small increase in the percentage of students who could recognize different graphical representations of the same data, although this was demonstrated by only slightly more than half of the students on the posttest. 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 to 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

7.4. BOXPLOTS

Students seemed to have many difficulties understanding and interpreting boxplots. A small percentage 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. Similarly, students did not associate a larger interquartile range with a larger standard deviation, given two boxplots with about the same range. There was no apparent gain in students' understanding that boxplots provide only estimates of percentages at the quartiles.

7.5. 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.

7.6. 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, although statistically significant, only a small percentage of students showed gains 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. About three fourths of the students did not demonstrate this understanding on the posttest. 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.

7.7. 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.

7.8. SAMPLING VARIABILITY

Students demonstrated difficulty with understanding sampling variability and sampling distributions. There was only a small increase in the percentage of students who demonstrated an understanding that statistics from relatively small samples vary more than statistics from larger samples, an understanding of expected patterns in sampling variability, or an understanding of factors that allow generalization from a sample to a population. Similarly, only a small percentage showed gains on an item that had them select a histogram representing a sampling distribution from a given population for a particular sample size. One of the most difficult items expected them to use sampling error as an appropriate measure when making an informal inference about a sample mean.

7.9. CONFIDENCE INTERVALS

Students did not demonstrate an understanding of confidence intervals. Whereas 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. About two thirds of the students understood that a confidence level does not represent the percentage of population values between the confidence limits. There was an increase in the percentage of students who incorrectly indicated that a confidence level represents the expected percentage of sample values between the confidence limits. The majority of students on the posttest also incorrectly indicated that a confidence level indicated the percentage of all sample means that fall between the confidence limits.

7.10. TESTS OF SIGNIFICANCE

Many students entered the course already recognizing that lack of statistical significance does not mean no effect. Most students indicated on the posttest that a low p-value is required for statistical significance. A small percentage of students made gains in identifying a correct interpretation of a significance test when the null hypothesis is rejected, although almost half did not demonstrate this understanding on the posttest. However, although a little over half of the students recognized a correct interpretation of a p-value, the majority of these students also responded that an incorrect interpretation was valid, indicating that many students hold both types of interpretation without recognizing the contradiction.

8. SUMMARY

The CAOS test provides valuable information on what students appear to learn and understand after 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 was 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. Instructors and researchers can register to use the CAOS test at the ARTIST website (https://app.gen.umn.edu/artist/). Plans are currently underway for the development of a collaborative effort among many institutions to gather large amounts of test data (including CAOS) and instructional data online 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.

ACKNOWLEDGMENT

The Research presented in this manuscript was supported by the National Science Foundation (NSF CCLI ASA-0206571).

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.

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.

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.

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.

delMas, R. & 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).
[Online: 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.
[Online: 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.
[Online: 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*. NSF CCLI grant ASA- 0206571.
[Online: https://app.gen.umn.edu/artist/]

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).
[Online: http://www.amstat.org/publications/jse/v3n1/konold.html]

Konold, C., & Higgins, T. (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., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 34*, 392-414.

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

Mathews, D., & Clark, J. (1997, March). 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.
[Online: http://fehps.une.edu.au/serj]

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

Pedhazur, E. J., and Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Erlbaum.

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.

Reed-Rhoads, T., Murphy, T. J., & Terry, R. (2006). *The Statistics Concept Inventory (SCI)*.
[Online: http://coecs.ou.edu/sci/]

Rubin, A., Bruce, B., & Tenney, Y. (1990, August). Learning about sampling: Trouble at the core of statistics. Paper presented at the Third International Conference on Teaching Statistics, Dunedin, 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.

Shaughnessy, M. (1977). 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.

Shaughnessy, 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). New York: MacMillan Publishing Company

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999, April). School mathematics students' acknowledgment of statistical variation. For the NCTM Research Presession 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.

ROBERT DELMAS
157 Education Sciences Building
56 East River Road
University of Minnesota
Minneapolis, MN 55455-0364
USA

APPENDIX A: PERCENT OF STUDENTS WITH ITEM RESPONSE PATTERNS FOR SELECTED CAOS ITEMS

			Item Response Pattern[a]			
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.	760	8.6	17.9	20.4	53.2
2	Ability to recognize two different graphical representations of the same data (boxplot and histogram).	759	26.0	17.8	28.6	27.7
3	Ability to visualize and match a histogram to a description of a variable (neg. skewed distribution for scores on an easy quiz).	760	20.8	6.1	22.5	50.7
4	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).	757	26.6	10.3	25.5	37.6
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).	758	23.0	5.9	21.1	50.0
6	Understanding to properly describe the distribution of a quantitative variable, need a graph like a histogram that places the variable along the horizontal axis and frequency along the vertical axis.	754	68.0	6.8	16.8	8.4
7	Understanding of the purpose of randomization in an experiment.	754	81.2	6.5	10.3	2.0
8	Ability to determine which of two boxplots represents a larger standard deviation.	755	21.3	19.5	24.0	35.2
9	Understanding that boxplots do not provide estimates for percentages of data above or below values except for the quartiles.	751	59.7	13.7	17.0	9.6

[a]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.

Item	Measured Learning Outcome	*n*	Item Response Pattern[a]			
			Incorrect	Decrease	Increase	Pre & Post
10	Understanding of the interpretation of a median in the context of boxplots.	754	62.3	9.4	18.0	10.2
11	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.	756	3.8	7.9	8.2	80.0
12	Ability to compare groups by comparing differences in averages.	753	4.8	9.4	10.0	75.8
13	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.	752	15.4	11.0	22.7	50.8
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.	746	38.6	9.7	27.1	24.7
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.	747	37.1	16.1	24.6	22.2
16	Understanding that statistics from small samples vary more than statistics from large samples.	747	60.2	7.9	17.0	14.9
17	Understanding of expected patterns in sampling variability.	746	37.3	12.5	20.0	30.3
18	Understanding of the meaning of variability in the context of repeated measurements and in a context where small variability is desired.	746	7.6	11.8	11.8	68.8
19	Understanding that low p-values are desirable in research studies.	730	21.1	10.4	32.7	35.8
20	Ability to match a scatterplot to a verbal description of a bivariate relationship.	748	1.9	5.6	7.6	84.9

Item	Measured Learning Outcome	*n*	Item Response Pattern[a]			
			Incorrect	Decrease	Increase	Pre & Post
21	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).	749	7.1	9.2	19.4	64.4
22	Understanding that correlation does not imply causation.	743	27.5	19.9	17.9	34.7
23	Understanding that no statistical significance does not guarantee that there is no effect.	735	17.6	18.1	19.3	45.0
24	Understanding that an experimental design with random assignment supports causal inference.	731	20.1	20.4	21.3	38.2
25	Ability to recognize a correct interpretation of a p-value.	712	23.5	22.1	29.8	24.7
26	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).	719	19.5	22.0	27.4	31.2
27	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is effective).	717	28.5	18.8	29.3	23.4
28	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits)	729	33.5	23.3	18.1	25.1
29	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).	725	24.4	8.0	43.0	24.6
30	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits)	723	38.9	16.9	29.7	14.5
31	Ability to correctly interpret a confidence interval.	720	16.0	9.7	36.9	37.4
32	Understanding of how sampling error is used to make an informal inference about a sample mean.	718	70.2	12.7	13.0	4.2

[a]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.

Item	Measured Learning Outcome	*n*	Item Response Pattern[a]			
			Incorrect	Increase	Decrease	Pre & Post
33	Understanding that a distribution with the median larger than mean is most likely skewed to the left.	730	36.6	23.7	21.9	17.8
34	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.	724	19.1	15.7	25.7	39.5
35	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.	719	39.4	16.4	26.1	18.1
36	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.	719	25.7	21.3	21.6	31.4
37	Understanding of how to simulate data to find the probability of an observed value.	722	67.3	13.2	12.3	7.2
38	Understanding of the factors that allow a sample of data to be generalized to the population.	715	50.5	11.6	23.5	14.4
39	Understanding of when it is not wise to extrapolate using a regression model.	710	63.9	11.5	18.2	6.3
40	Understanding of the logic of a significance test when the null hypothesis is rejected.	716	31.6	16.5	26.5	25.4

[a]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 MISUNDERSTANDING AND MISCONCEPTIONS

			Item Response Pattern[a]			
Item	Misconception or Misunderstanding	*n*	Neither	Decrease	Increase	Pre & Post
6	A bell-shaped bar graph to represent the distribution for a quantitative variable.	754	31.6	15.6	25.5	27.3
7	Random assignment is confused with random sampling or thinks that random assignment reduces sampling error.	754	36.5	14.3	27.3	21.9
15	When comparing histograms, the graph with the largest number of different values has the larger standard deviation (spread not considered).	747	52.9	14.1	20.6	12.4
22	Causation can be inferred from correlation.	743	50.9	13.2	22.1	13.9
36	Grand totals are used to calculate conditional probabilities.	719	51.3	15.3	23.5	9.9
40	Rejecting the null hypothesis means that the null hypothesis is definitely false.	716	50.4	17.2	22.9	9.5

[a]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.