STUDENTS'MISCONCEPTIONS ABOUT INTRODUCTORY STATISTICS TOPICS:

ASSESSING STAT 2000 OUTCOMES USING CAOS

by

ELIZABETH BARBARA AMICK

(Under the Direction of Jennifer Kaplan)

ABSTRACT

Assessments play an important role within any subject as a method to determine how well students are learning. Until recently within the field of statistics education there has been little attention paid to evaluating assessments. In order to improve this, an assessment known as the CAOS was developed and it has been found to be a valid and reliable assessment for introductory statistic students. In this study, the CAOS test will be examined and the scores of the CAOS test from the University of Georgia STAT 2000 students will be analyzed. Results are reported on what statistical concepts students understand and what concepts students are struggling with the most. The UGA scores are then compared to the national CAOS scores to determine any similarities and differences.

INDEX WORDS:	Statistics education research; Assessment; CAOS; Introductory statistics; Misconceptions

STUDENTS'MISCONCEPTIONS ABOUT INTRODUCTORY STATISTICS TOPICS:
ASSESSING STAT 2000 OUTCOMES USING CAOS

by

ELIZABETH BARBARA AMICK

B.S., University of South Carolina, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

STUDENTS'MISCONCEPTIONS ABOUT INTRODUCTORY STATISTICS TOPICS:

ASSESSING STAT 2000 OUTCOMES USING CAOS

by

ELIZABETH BARBARA AMICK

Major Professor:	Jennifer Kaplan
Committee:	Jack Morse
	Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

DEDICATION

I would like to dedicate my thesis to my wonderful mother, brother, and father. My mother has always been there for me throughout everything with her never-ending love and support. She is my number one supporter and I don’t know what I would do without her. Although I am the older sister, I believe I am the one who looks up to my younger brother. I hope to one day possess his amazing generosity, genuine kindness, and brilliant sense of humor. He is the one person who makes me laugh like no one else. I am so proud that he is my brother. Last, I want to thank my father. He is someone I have always admired. He is the reason I have made it this far in my life. Without my father’s advice and teachings I am not sure I would be where I am today. Not only is he the best statistician I know, he is the best dad in the world. I am so blessed and thankful to have such an amazing family by my side.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank my amazing advisor Dr. Kaplan. Her enthusiasm and passion for statistics education has made the entire thesis process enjoyable. I would also like to thank my thesis group: Greg, Adam, and Kristi. Their opinions and advice helped me immensely and I could not have asked for a better group!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In order to determine how much a student has learned in a specific course, end of the semester tests are administered. It is important to assess what knowledge a student has gained from taking the course. Bingham (2001) argues that assessment is an integral part of the whole process of teaching and learning. Assessments allow students and teachers to evaluate their knowledge base, give teachers insight into their own effectiveness, allow institutions to award grades to the students, and provide society a way to judge the effectiveness of the institution (Davies & Marriott, 2010). It is through assessment that students learn about their individual strengths and weaknesses and current level of skills and understanding (ibid). Starkings describes a key role of assessment as the "diagnostic process- by establishing what students have learned, it is possible to plan what students need to learn in the future" (1997, pg. 139).

Statistics can be considered a very broad and diverse subject, since it can be applied to a variety of other subjects, including the sciences, business, economics, geography and psychology (Davies & Marriott, 2010). Due to the fact that statistics is a broad subject with many components, it is often difficult to design effective assessments (MacGillivray, 2010). Gal and Garfield state, "Educators are further challenged by the need to make sure that students understand the real-world problems that motivate statistical work and investigations" (1997, pg. 5). Furthermore, Garfield (2002) reported assessment practices in statistics education have undergone the least amount of reform when considering the body of reforms in statistics education in the last two decades.

In this thesis, the Comprehensive Assessment of Outcomes in Statistics, (CAOS) assessment was examined. The CAOS assessment is a confirmed valid and reliable assessment consisting of multiple-choice items that cover concepts taught in a first statistics course. The items on the CAOS require students to have a more conceptual understanding of the topics instead of the ability to apply straightforward calculations or use formulas. The data collected from the CAOS assessment is used to answer the question of what students know after completing an introductory statistics course.

In the fall 2012 semester, the CAOS assessment was administered to UGA STAT 2000 students at the end of the semester. The goal of the research is to determine the areas of strength and weakness for STAT 2000 students by observing their CAOS scores. In addition, the UGA STAT 2000 students' scores will be compared with the national CAOS scores in order to determine the similarities and differences between UGA students and statistics students nationwide.

Chapter 2 presents a literature review of the common statistical misconceptions determined from the national CAOS study. Chapter 3 describes in detail the development of the CAOS assessment and how the creators of the CAOS assessed its validity and reliability. In Chapter 4, we begin with a description of STAT 2000 and follow with the data collection and analysis methods. In Chapter 5, STAT 2000 students' scores will be analyzed to determine the areas of strength and weakness and the STAT 2000 scores will be compared to the national CAOS scores. Chapter 6 concludes the thesis with a summary and recommendations for the future of UGA STAT 2000.

CHAPTER 2

LITERATURE REVIEW OF CAOS CONCEPTS

The authors of the CAOS claim that the test examines ten statistical concepts: descriptive statistics, bivariate data, graphical representations, boxplots, data collection and design, probability, sampling variability, significant tests, confidence intervals, and the normal distribution. The analysis of content of the CAOS done for this study and reported in Chapter 4 did not indicate any items on the CAOS that assess directly student understanding of the concept of the normal distribution. This chapter, therefore, does not include literature results on student understanding of the normal distribution. The literature on the other nine concepts tested by the CAOS is presented. The concepts are presented in the order specified above.

2.1 Misconceptions about Descriptive Statistics

The one concept associated with the topic of descriptive statistics found to be most misunderstood in the national results of the CAOS assessment was variability. Statistics has been described as the science of variation by MacGillivray (2004). In addition, consideration of variation has been proposed as one of the fundamental types of statistical thinking (Wild & Pfannkuch, 1999) or as a major factor contributing to the development of students' statistical thinking (Meletiou-Mavrotheris & Lee, 2002). Bakker stated that students who do not expect variability will lack "intuition of why one would take a sample or look at a distribution" (2003, pg. 3). The terms *variability* and *variation* are both used in the literature describing student understanding of descriptive statistics. The term *variability* can be taken to mean the characteristic of the entity that is observable. In other words, variability refers to a set of measures, such as interquartile range or standard deviation.

Reasoning about variation can be broken down into four components as described by Wild and Pfannkuch (1999). The four components of Wild and Pfannkuch's (1999) consideration of variation are:

1. Noticing and acknowledging variation
2. Measuring and modeling variation for the purposes of prediction, explanation, or control
3. Explaining and dealing with variation
4. Using investigative strategies in relation to variation

These components provide the groundwork for expanding on the notion of understanding variation. Despite the central role variation plays in statistics (Hoerl & Snee, 2001) there is very little research on a student's understanding of the topic (Reading & Shaughnessy, 2004). Garfield, delMas and Chance (1999) found that students presented with a histogram judged the variability of the distribution on the basis of the variation in the heights of the bars, instead of the relative density of the data around the mean. It is important to help students develop a better understanding of variability and its representation in order to help support a better understanding of sampling distributions (delMas & Liu, 2005).

2.2 Misconceptions about Bivariate Data

Estepa and Cobo argue "association has great relevance for the training of researchers since it is essential for many statistical methods and techniques frequently used by researchers" (2001, pg. 37). In addition they state that difficulty in understanding association and topics related to association can result in "misinterpretations and misuses of statistics methods in research" (ibid, pg. 37). Batanero and Estepa (1996) investigated 18-year-old students' understanding of correlation between numerical variables presented in a scatter plot. This type of error is not limited to analysis of bivariate data and is linked to understanding of experimental design, which will be discussed in Section 2.5.

Another issue with student understanding of bivariate data, discovered by Morris (1997), is that some students accredit positive correlation to a stronger association than negative correlation. In addition, the students believed that a negative correlation signifies the variables are independent. Brousseau (1997) found that when calculating the rank of the strength of correlations, many students had difficulties properly ordering the negative correlations. Batanero, Godino, and Estepa (1998) found similar results, as students did not use the greatest absolute value to represent the greatest correlation. In an investigation done by Batanero and Estepa (1996), they found that students use only part of the data provided in a scatterplot when making judgments from the data. If the partial information confirmed a specific type of correlation the students used this association in their answer. Estapa and Cobo (2001) conducted a study of association with 193 undergraduate students who had finished an introductory statistics course and found that concepts related to association such as: correlation, covariance, and regression were only understood by approximately 53.2 % of the students.   The authors use the poor results to argue for the need for emphasizing these concepts in the teaching of association.

2.3 Misconceptions about Graphical Representations

Graphs are important for data representation, data reduction, and data analysis in statistical thinking and reasoning (Shaughnessy, 2007). Friel, Curcio, and Bright define graph understanding as “the ability of graph readers to derive meaning from graphs created by others or by themselves” (2001, pg.132). Traditionally, statistics instruction focuses mainly on the construction of graphs, leaving out more conceptual understandings such as interpreting or making predictions based on the graphical information (Friel et al., 2001). Maxine Pfannkuch argues that graphs are frequently used as "illustrations of data rather than as reasoning tools" (2006, pg. 27, see also Wild and Pfannkuch, 1999). Recent research in statistics education has shown students have difficulty understanding graphical representations of distributions (e.g., Bakker & Gravemeijer, 2004 2004; Ben-Zvi 2004; Biehler, 1997; Hammerman & Rubin, 2004; Konold, 2003; McClain, Cobb, &Gravemeijer, 2000). Results from these recent studies indicate that

1. Students tend to focus on individual points of a data set or graph such as outliers rather than observing a graph or a data set in its entirety.
2. Understanding that area on a histogram represents a measure of frequency is not an intuitive notion.
3. Understanding data as an entity in a graph involves coordinating ideas of center, variability, density and skewness.
4. Students are most familiar with bar graphs or case value graphs, where each case or data point is represented by a bar or a line, and the ordering of these is arbitrary, and are not as familiar with graphs in which distributions of quantitative variables are shown in aggregate, such as in a histogram.
5. Even when making comparisons of distributions, novices tend to compare slices of data or points, rather than comparing entire entities, taking into consideration overall center and spread.

Curcio (1989) identified three levels of graphical understanding that students should attain to master graphical literacy: reading the graph, reading within the graph, and reading beyond the graph. In order for a student to be "reading" a graph he or she must be able to understand the graph's scale and measurement units. Students who read "within" the graph are able to interpret any graphical trends and patterns. Reading "beyond" the graph, the highest level of Curcio's graph interpretation involves a student who has the ability to ask questions about the dataset and can also project into the future. For instance, a student may ask where the data came from or how they were collected. In addition, if a graph shows a decreasing pattern over time, a student would be able to infer that the data may level off sometime in the future (Shaughnessy, 2007). These three levels contribute to the ability of a student to read graphs.

A summary analysis of Friel et al. (2001) provides six behaviors that they considered to be closely associated with graph sense. These behaviors can be connected to Curcio's three levels of graph comprehension.

1. Recognizing components of graphs (Reading the data).
2. Speaking the language of graphs (Reading the data).
3. Understanding relationships among tables, graphs, and data (Reading within the data).
4. Making sense of a graph, but avoiding personalization and maintaining an objective stance while talking about the graphs (Reading within the data).
5. Interpreting information in a graph and answering questions about it (Reading beyond the data).
6. Recognizing appropriate graphs for a given data set and its context (Reading beyond the data).

2.4 Misconceptions about Boxplots

A challenge for students is to understand the nature and type of reasoning involved when making informal inferences from sample distributions about population distributions. Boxplots are graphical tools that can help students make informal inferences. Boxplots condense, summarize, and obscure information, and incorporate statistical notions such as median and quartiles (Bakker, 2004). Basic boxplots are introduced to students from as young as 12 years of age in the United States (Bakker, Biehler, & Konold, 2005). Statistic instruction on the topic of boxplots traditionally involves the basic construction of the boxplot and nothing more. This results in students not knowing the reasoning tools behind the construction of the boxplots. (Friel et al., 2001).

Informal inferential reasoning is being able to infer that one group is generally greater than a second group, or that no distinction can be drawn, by observing the boxplot distributions (Pfannkuch, 2006). If a student possesses boxplot-reasoning tools, then he or she will be able to correctly draw informal inferences.   The question arises as to what elements of reasoning are necessary for comparing boxplot distributions (Pfannkuch, 2006). Formal inferential reasoning concentrates on the centers of the distributions as being representative of the data. Therefore, an important element in developing the reasoning process about inference is for students to be able to incorporate ideas about the middle part of the data as a way to characterize the data (Pfannkuch, 2006). The middle part of the data is shown in the box portion of a boxplot. In addition, a measure of variability must be included in a student’s reasoning process. The length of the box in the boxplot provides a measure of variability, the interquartile range. This reasoning must include notions of comparing variability, or the lengths of the boxes, within and between boxplots. With these reasoning tools present, students will be able to make informal inferences about the population. Clearly, the ability to reason informally about inference is linked not only to the concept of boxplots, but also to the concept of variability, discussed in Section 2.1.

2.5 Misconceptions about Data Collection and Design

There is little published research on student learning of experiment design, such as sampling techniques and survey or study design. In one paper that was found, Perrett (2012) argues that the concept of experimental units is a key topic for statistical education. Furthermore, he claims it is a concept that tends to be difficult for students to understand. The experimental unit could be defined as a group of individuals or just a single individual. In addition, the experimental unit is determined by how the treatments are assigned (Perrett, 2012). For example, consider an experiment investigating a treatment on a group of mice with two possible design scenarios. Scenario A involves having all the mice in one cage receiving the same treatment while scenario B involves the each mouse being separated into its own cage and given the same treatment. For scenario A the experimental unit is the group of mice because the treatment is non-independent as the mice were together when receiving the treatment. In scenario B the experimental unit is each individual mouse because each mouse is independent of other mice and the treatment is considered independent. Thus experimental unit can be more specifically defined as the object independently treated in an experiment. The word independent is a key component of an experiment aimed at proving cause and effect such as the example with the mice. If the experimenter wanted to be able to link the treatment on the mice to the cause of a certain effect then he or she must be sure to have correctly identified the experimental unit.

Another issue that arises when the experimental unit is incorrectly identified is that Type-1 error rates in hypothesis testing can become inflated (Blair, 1983). In reviewing medical journals it was discovered that the rate at which experimental unit was not correctly identified was 44% (Calhoun, Guyatt, Cabana, Lu, Turner, Valentine, & Randolph, 2008). The editorial board of the *Journal of Teaching in Physical Education* suggested appropriate ways of correcting the errors by describing the experimental unit in terms of topics such as treatments, random assignment etc. (Silverman & Solmon, 1998). If a student does not understand the term experiment unit it can lead to several other misunderstandings. For instance, a student who does not know how to identify an experimental unit will not be able to calculate the degrees of freedom for an experiment as degrees of freedom depend on knowing the number of experimental units within certain conditions. Perrett (2012) researched how the term experimental unit was defined in statistical textbooks and students' responses to the Advanced Placement Statistics exam. Perrett concluded the way the concept of experimental unit was defined had an impact on how a student performed.

As mentioned previously, there is little published research in this area. There is anecdotal evidence, however, that students tend to prefer to use a representative, (non-random convenience) sample selected based on some known factors over a random sample for making inferences about the population (Milo Schield, personal communication). It took some time to convince these students that the benefit of random sample, in terms of getting a representative sample on previously unknown factors was extremely important. After a lengthy discussion, the students indicated a preference for the using stratified random sampling since that seemed to be representative on both known and unknown factors.

The misconceptions discussed in the section on bivariate data in which students confuse association with causation are related to misconceptions in data collection and design, in that causal inference is typically only done when a study uses a randomized experimental design. Textbooks, however, are very uneven in talking about causal inference and design. They may never use the word "confounding" or "lurking" variable and different textbooks provide different definitions for the two terms. In some cases, textbooks use causally related words, "effect", “result” and various action verbs (“influences”, “reduced”, “vary with a change”, “improves”, etc.). But all too often they finish by saying “experiments establish relationships”; “experiments do not establish causal relationships” (Hinton, 1995, pg. 75). In other cases, textbooks move from correlation to causation via random assignment. Experiments are conducted to view whether “the variables are linked in a cause-and-effect manner” –“to see whether the independent variables affect the dependent variable” (Huck & Cormier, 1996, pg. 584-586). The term “affect” can have many synonyms including: influence, determines, creates, improve, etc. If these terms are used within a statement concerning the independent and dependent variables, the focus of the study was more than likely cause-and-effect. The process of randomly assigning subjects is “a defining characteristic of experiments” (Huck & Cormier, 1996, pg. 584-586). The random allocation method, when used in a properly conducted experiment, will provide impartial assignment of extraneous influences among the groups being compared. Therefore, the differences among the observed group are caused solely by the differences in treatments the groups received and not on how the groups were assigned (Everitt, 1996).

2.6 Misconceptions about Probability

Halpern states, “Probability is the study of likelihood and uncertainty. It plays a critical role in all of the professions and in most everyday decisions” (1996, pg. 242). The National Council of Teachers of Mathematics (NCTM) (1991) recognized the importance of being able to reason effectively about probability. It was recommended by the NCTM that students be capable of reasoning about probability and drawing inferences. Hirsch and O'Donnell state, “Unfortunately, current secondary school curricula are only beginning to incorporate statistical skills, and, as a consequence, most students enter college with very little formal experience with the laws of probability and probabilistic reasoning” (2001).

Misconceptions when reasoning about probability can occur because of violations in the application of laws of probability (Hirsch & O’Donnell, 2001). Examples of such errors include stereotyping, confirmation bias, and matching bias. Students often form misconceptions through informal experiences outside the classroom (Garfield & Ahlgren, 1988). In addition, students may develop their own way of reasoning about uncertain events (Kahneman & Tversky, 1972). Their lack of understanding may be due to a lack of experience with the mathematical laws of probability or because they use heuristics. Even students who receive formal instruction continue to have misconceptions about the nature of probability and probabilistic reasoning (Kahneman & Tversky, 1972). Shaughnessy (1992) stressed the need to (a) know more about how students think about probability, (b) identify effective methods of instruction, and (c) develop consistent, reliable methods of assessment that more accurately reflect students' conceptual understanding. There is a need to develop consistent and reliable methods for accurately assessing students' conceptual understanding of probability, in order to evaluate instructional methods (Hirsch & O’Donnell, 2001).

2.7 Misconceptions about Sampling Variability

Sampling distributions are central to statistical inference (Castro Sotos, Vanhoof, Noortgate & Onghena, 2007). Although many students are able to understand the sampling process alone, they often cannot properly use concepts in inferential reasoning (Batanero, 2005). Understanding of sampling distribution is a foundation for understanding of how to test hypotheses and construct confidence intervals. The main idea in inferential statistics is that a sample provides some but not all information about the population from which the sample was drawn (Castro et al., 2007). Students need to be able to connect the population to the sample(s)

(Shaughnessy, 2007) in order to make inferences about the population based on the information contained in the sample.

Work done by Saldanha and Thompson (2003) provides evidence that their students did not have a sense of variability that extended to ideas of distribution. This is perhaps not surprising given the information presented in Section 2.1 on student misconceptions of variability. Rubin, Bruce and Tenney (1991) describe peoples' understanding of conclusions that can be drawn from a sample as a range from "knowing everything about a sample" to "knowing nothing about a sample". "Knowing everything" corresponds to the belief that samples should be perfectly representative of the population. "Knowing nothing" is the belief that a population can never be represented by a sample. This belief occurs when people are preoccupied with the idea of variability and believe that a sample is just chance. For students to have a solid understanding of sampling variability there needs to be a balance between the "knowing everything" extreme and the "knowing nothing" extreme (Watson & Kelly, 2004).

Many misconceptions about the sampling process relate to the sample mean (Castro Sotos et al., 2007). It has been shown that many students do not understand the effect of sample size on the variance of the sample mean (Chance, delMas, & Garfield, 2004). Tversky and Kahneman (1971) proposed the idea of *representativeness heuristic* which states that people believe samples behave similarly to the population, regardless of the sample size. This is similar to Rubin, Bruce and Tenney's extreme of "knowing everything" mentioned above. If a student has this belief he or she will not be able to make valid inferential conclusions of generalizations to the population (Innabi, 1990).

2.8 Misconceptions about Significance Tests

An important topic in statistics is the significance test (also called a hypothesis test). In a significance test we evaluate the evidence from a sample against a previously defined null hypothesis. Statistics instructors often have a difficult time helping students understand significance testing (Kirk, 2001). The main reason for this is that performing a significance test requires the understanding of abstract concepts such as the p-value, sampling distribution, and the significance level. There are many levels and components of hypothesis testing that students do not understand (Castro Sotos et al., 2007). One such component is knowing the difference between the alternative and null hypotheses. Specifying the null and alternative hypotheses is one of the first steps of significance testing and if done incorrectly will result in false conclusions. Vallecillos and Batanero (1996) remarked that students can have issues identifying and defining the null and alternative hypothesis. The two specific misconceptions concerning hypotheses found in Vallecillos and Batanero (1997) were:

1. Believing that the null hypothesis and the acceptance region “are the same thing”
2. Believing that a hypothesis can refer both to a population and a sample

Another misconception about hypothesis testing deals with the significance level. Significance level along with the p-value can be considered the most complicated concepts of significance testing (Haller & Krauss, 2002). The definition of the significance level is the probability of rejecting the null hypothesis when it is true and can be seen in notation form below:

$$\alpha = P\text{ (Rejecting } H_o \mid H_o \text{ true)}$$

Falk (1986) argues the most common misconception of the significance level is the switch of the two terms in the conditional probabilities. When this occurs the definition of the significance level becomes defined as the probability that the null hypothesis is true once the decision to reject it has been taken, meaning:

$$\alpha = P\ (H_o \text{ is true once we have rejected } H_o)$$

This misconception can be seen in a study done by Vallecillos (2002). She found that 53% of the 436 university students believed an item stating, "A level of significance of 5% means that, on average, 5 out of every 100 times that we reject the null hypothesis, we will be wrong".

In addition to the significance level, there are misconceptions concerning the concept of a p-value. One misconception is students consider a p-value's numeric value as an indicator of the strength of the treatment effect (Castro Sotos et al., 2007). Gilner (2002) states that lower p-values are sometimes understood by students as having stronger treatment effects than those with higher p-values. Lane-Getaz (2008) evaluated the Reasoning about P-values and Statistical Significance (RPASS) scale with a sample of 177 students from seven introductory and intermediate statistics and probability courses. It was shown that respondents had difficulty understanding that the p-value depends on the alternative hypothesis. Similarly, respondents had difficulty understanding the sample size impact on statistical significance.

2.9 Misconceptions about Confidence Intervals

As the research on the misconceptions of significance testing grows, many educators have promoted a wider use of confidence intervals (Cumming, Williams, & Fidler, 2004; Harlow, Mulaik & Steiger, 1997). The American Psychological Association supports confidence intervals as being the best statistical reporting strategy (APA, 2001). Furthermore, when learning about hypothesis testing, confidence intervals can help decrease the misconception surrounding statistical significance (Fidler, 2006). Nevertheless, confidence intervals are sometimes incorrectly interpreted. The results of a study done by Belia, Fidler, Williams, and Cumming

(2005) indicated many researchers have fundamental and severe misconceptions about how confidence intervals can be used to support inferences from data. Of their 473 respondents, the results showed that many leading researchers have misconceptions about how error bars relate to statistical significance. They also showed that many researchers could not adequately distinguish the confidence interval and standard error bars.

Along with researchers, students have been shown to be prone to confidence interval misconceptions. Fidler (2006) performed a series of experiments with 180 psychology and ecology students and found several misconceptions dealing with confidence intervals. A few of the higher frequency misconceptions include:

1. Belief that an interval with a lower confidence interval is wider than an interval with a higher confidence interval(for the same data)
2. Thinking that the interval is a set of plausible values for the sample mean
3. The width of a confidence interval is not affected by sample size

In this chapter we have summarized the literature on the misconceptions associated with nine of the statistical concepts assessed by the CAOS: descriptive statistics, bivariate data, graphical representations, boxplots, data collection and design, probability, sampling variability, significant tests, and confidence intervals. These concepts are all taught as part of the STAT 2000 course at UGA so the results of the literature should inform the teaching of STAT 2000 in the future.   In the next chapter we begin to examine how the CAOS assessment was developed.

CHAPTER 3

THE CAOS ASSESSMENT

The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project is a National Science Foundation (NSF) funded project that addresses the assessment challenges facing statistics education as outlined by Garfield and Gal (1999). Garfield and Gal describe the need for reliable, valid, and practical assessment items for statistics education. The ARTIST project developed an overall Comprehensive Assessment of Outcomes in Statistics (CAOS). The CAOS test was designed to be a reliable assessment consisting of items important to an introductory statistics course.

The CAOS test was developed through a three-year process. The process included obtaining existing items from instructors, revising these items, collaborating with advisors and class testers, and writing any additional items not covered. In addition, two large content validity assessments were performed. The ARTIST advisory board provided the validity ratings of items. The validity ratings determined the content validity for the targeted population of students (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). As the ARTIST advisory board was reviewing the items it was decided that main focus of the CAOS test should be devoted to different facets of reasoning about variability. The concept of variability was viewed as the primary goal for an introductory statistics course. Variability was extended to cover variability in distributions, comparing groups, sampling and sampling distributions.

The initial items created for the CAOS assessment were chosen from the ARTIST online database consisting of over 1000 multiple-choice items. Each item was reviewed to ensure the items followed established guidelines for multiple choice items (Haladyna, Downing & Rodriguez, 2002). Each item was evaluated by the ARTIST advisory team to determine the content validity and discover any topics that may not have been covered by the test. After feedback and several revisions the ARTIST team created the first version of the CAOS test (CAOS 1), which consisted of 34 multiple-choice items. These items were written to make students think and reason as opposed to computing or recalling definitions and formulas.

CAOS 1 was used in a pilot study with introductory statistics students during the fall 2004. The pilot study provided the data to make revisions to CAOS 1, leading to the second version of CAOS (CAOS 2), which consisted of 37 multiple-choice items. The CAOS 2 test was administered to nearly 100 secondary-level students and 800 college-level students. The results from CAOS 2 were used to create a third version of CAOS (CAOS 3). CAOS 3 was given to 30 statistics instructors who were faculty graders of the Advanced Placement Statistics exam in June 2005. The instructors reviewed CAOS 3 to determine the validity of the items. It was determined by the instructors' ratings that CAOS 3 was measuring what it was designed to measure, but the instructors contributed several suggestions for additional changes. Their feedback was used to produce a final version of the test called CAOS 4, which consisted of 40 multiple-choice items.

In March 2006 there was a final discussion on the content validity of CAOS 4. Each item on the CAOS 4 was reviewed by 18 members of the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE). These individuals teach statistics at the college level and are considered experts in the national statistics education community. Each individual was given a copy of CAOS 4 with an outline of what each question was designed to measure. After reviewing the test, the members were asked questions on the validity of the items. There was unanimous agreement by 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 year courses in statistics" (delMas, Garfield, Ooms, & Chance, 2007, pg. 31). Similarly, 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 course" (ibid). Although some raters did imply that some topics were missing there were no additional topics recognized by the majority of the raters. The evidence gained by the raters provided evidence that the assumption of content validity was met.

In the fall of 2005 and spring of 2006, CAOS 4 (henceforth referred to as CAOS) was administered as an online and hard copy test. The test was given to a total of 1470 introductory statistics students, taught by 35 instructors from 33 higher education institutions from 21 states across the United States. It was determined from the 40 items on the CAOS posttest that the estimated internal consistency reliability was well above the range of suggested lower limits with a Cronbach's alpha coefficient of 0.82. Of the 1470 students, 763 students were chosen to take a CAOS pretest and a posttest. The 763 students were taught by 22 instructors at 20 higher education institutions from 14 states in the United States.

The results from this national study showed that there was a small overall increase in correct response from pretest to posttest. The average percentage for correct responses on the CAOS pretest was 44.9% corresponding to the total number of correct points equaling 17.96 out of 40. The posttest showed a small increase of 9 percentage points with an average correct percentage response equaling 54% corresponding to the total number of correct points equaling 21.6 out of 40. There were three areas of items that could be separated by the results of the CAOS test: (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, (c) items that were more difficult for students to learn. There were ten concepts that students showed an increase in misconception from pretest to posttest. The ten concepts include: sampling variability, graphical representation, probability, boxplots, significant tests, confidence intervals, data collection and design, descriptive statistics, bivariate data, and normal distributions.

CHAPTER 4

DATA COLLECTION AND METHODS

There are approximately 1300 students enrolled in STAT 2000 each spring and fall semester. STAT 2000 is a 4-hour credit course with 3 hours devoted to lecture and 2 hours for weekly lab sessions. The online course description for STAT 2000 states,

> Introductory statistics including the collection of data, descriptive statistics, probability, and inference. Topics include sampling methods, experiments, numerical and graphical descriptive methods, correlation and regression, contingency tables, probability concepts and distributions, confidence intervals, and hypothesis testing for means and proportions.

The CAOS assessment was administered during the last lab class of the semester. The STAT 2000 coordinator Jack Morse gave this instruction to all STAT 2000 students,

> This week in lab, you will be completing a statistical assessment that assesses how well students understand the basic concepts taught in most introductory statistics courses. Everyone completing this assessment will receive a 100 for your lab 10 grade. In addition, in order to motivate you on this assignment, for every question answered correctly, you will receive a quarter of a point towards your overall test point total. There are 40 questions on this assessment, so you could receive up to a maximum of 10 points towards your test point total. All the questions on this assessment are conceptual multiple choice questions, and you will not need StatCrunch nor a calculator on any of the questions. You will have 45 minutes to complete the assignment and you cannot get help from the TA.

The data collected from the CAOS results included students' identification number (all names stripped), total number of correctly answered items, and students' answer to each of the 40 multiple choice questions.

Recall from previous chapters that the CAOS is reported to test ten concepts typically taught in an introductory statistics course. In order to determine which concept each item on the CAOS was measuring, the members of a statistics education research group at UGA were asked to classify each item by the concept it tested. The group consisted of two UGA statistics graduate students: Greg Jansen and Kristi Clark, Dr. Jennifer Kaplan, and math education Ph.D. candidate, Adam Molnar. Individually each person read each CAOS question decided which of the ten concepts corresponded to the question. A question could address multiple concepts. After each person identified the concept(s) associated with each question, the results were combined into one table. The contents of the table were investigated by the group to determine if there were any discrepancies. The questions that did not have a clear corresponding concept were discussed among the group. After the discussion everyone came to an agreement on a list of misconceptions for each question (see Table 1 below). Notice that some CAOS items have been listed under multiple concepts. For example, item 14 was classified as assessing knowledge of descriptive statistics and graphical representations.

Table 1: CAOS Concepts Assorted by Item

Statistical Concept	CAOS Item numbers
Descriptive Statistics	8 14 15 33
Bivariate	20 21 22 39
Graphical Representation	1 2 3 4 5 6 11 12 14 15 33 34 35
Boxplots	2 8 9 10
Data Collection and Design	7 13 22 24 38
Probability	17 36 37
Sampling Variability	16 17 18 32 34 35
Tests of Significance	19 23 24 25 26 27 40
Confidence Interval	28 29 30 31

Using the breakdown show in Table 1, each concept and its corresponding items were investigated in more detail. The difficulty of each item was determined, along with the discrimination of each item. Item discrimination describes how well a question differentiates between high-scorers and low- scorers. By determining this value one can determine the topics that separate the high- scorers from the low-scorers. In addition, it could be that no one topic discriminates more than other topics, which is also useful to know. In order to determine both the difficulty and discrimination of an item, the top 200 STAT 2000 students with the highest number of total correct responses and the lowest 200 persons with the lowest number of total correct responses were compared. In order to calculate the difficulty of each item the proportion of students (out of the 400) who answered the question correctly was calculated. Table 2 below shows the difficulty index table.

Table 2: Difficulty Index

% Answered Correctly	Question Difficulty
0.75-1.00	Easy
0.60-0.74	High Average
0.40-0.59	Average
0.26-0.39	Low Average
0-0.25	Hard

In order to calculate the discrimination for each question the number of correct responses from the low scoring group was subtracted from the number of correct responses from high scoring group. This number was divided by the total number of people in each group (200). This can be seen in the formula below,

$$Discrimination = \frac{\#\ Correct\ HSG}{n_{HSG}} - \frac{\#\ Correct\ LSG}{n_{LSG}}$$

Table 3 below displays the discrimination index table.

Table 3: Discrimination Index

Discrimination	Description/Meaning
0.60-1.00	Very strong discriminator
0.40-0.59	Strong discriminator
0.20-0.39	Moderate Discriminator
(-0.19)-0.19	Non-Discriminator
-1.00-(-0.20)	Strong negative discriminator

CHAPTER 5

RESULTS

5.1 STAT 2000 CAOS Results

The number of STAT 2000 students that participated in the CAOS assessment was 1202. Figure 1 and Table 4, below, show the distribution and summary statistics for the results. The distribution of scores is unimodal and approximately symmetric, with perhaps a bit more high scores than would be expected in normal distribution, but not enough to invalidate an assumption of normality or suggest that the data have any skew. The average number of questions answered correctly is approximately 21.7 out of 40, or roughly 54%, with a standard deviation of 4.08 questions. Notice that no students scored less than 9 correct or higher than 35 correct, again out of 40 total questions. These values will be compared to those reported in the national sample in Section 5.2.

Table 4: UGA Summary Statistics for Total Number Correct

Analysis Variable: TotalCorrect				
N	Mean	Std Dev	Minimum	Maximum
1202	21.74	4.08	9.00	35.00

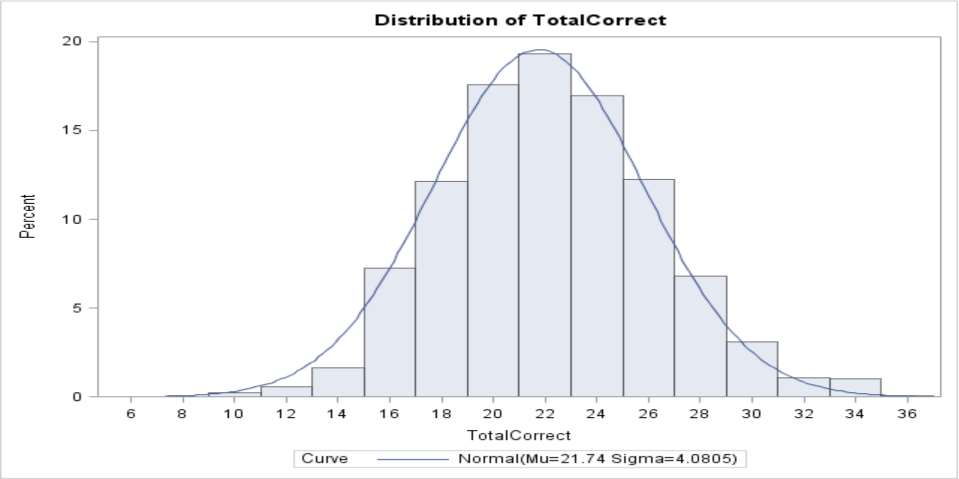

Figure 1: Histogram of Total Correct for UGA

The first concept that was analyzed was descriptive statistics. Table 5, below, shows the item numbers, the percent of students who answered the items correctly, the difficulty and discrimination values, and the measured outcomes for each item as described by the CAOS assessment. The lowest scoring item for descriptive statistics is item number 33 with 29.1% correct. The range for total percent correct is 40.5%, with item 8 having the highest percent correct at 69.6%. For the lowest scoring item, item 33, students were asked to choose which histogram best describe a set of summary statistics including the: mean, median, standard deviation, minimum and maximum. In order for a student to have gotten this question right, he or she would have to know that a distribution with a median larger than a mean is usually skewed to the left. This question can be considered harder than average on the difficulty scale and is a moderate discriminator. Item 33 is considered the hardest question in the group of items for descriptive statistics. For the highest item that was answered correctly, item 8, students were given two boxplots and asked which boxplot displays the larger standard deviation. This item is easier than average and has a moderate discrimination value.

Table 5: Descriptive Statistics

Descriptive Statistics				
Item#	%Correct	Difficulty	Discrimination	CAOS Measured Outcomes
33	29.11	Low Average	Moderate	Understanding that a distribution with the median larger than mean is most likely skewed to the left.
14	58.75	High Average	Strong	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.
15	59.6	High Average	Moderate	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.
8	69.58	High Average	Moderate	Ability to determine which of two boxplots represents a larger standard deviation.

The second concept was bivariate data as shown below in Table 6. The range for percent correct is 88%, which is a very large range indicating there was a very high scoring item and there was also a very low scoring item. The item that was answered correctly the least is item 39 with 8.6%. Item 39 is categorized as a hard question and is the hardest question in bivariate data and does not discriminate between higher and lower scorers. For item 39, students were required to recognize that it is not appropriate to extrapolate a regression model to values of the predictor variable that are well beyond the range of values in the study. The highest percent that was answered correctly corresponds to item 20. Item 20 is categorized as an easy item with no discrimination present. For item 20, students were asked to identity a scatterplot to the description of a bivariate relationship.

Table 6: Bivariate Data

Bivariate				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
39	8.63	Hard	None	Understanding of when it is not wise to extrapolate using a regression model.
22	47.62	Average	Strong	Understanding that correlation does not imply causation.
21	82.08	Easy	Moderate	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).
20	96.42	Easy	None	Ability to match a scatterplot to a verbal description of a bivariate relationship.

The third concept analyzed was graphical representation and is shown in Table 7. This concept is tested by the largest number of items of any of the 10 concepts, as graphical representation covers a wide area of topics. The range for percent of correct responses for this concept is 76.9%. The item that was the hardest for students to understand was item 6 with a percentage of total correct equaling 17.5%. Item 6 has a difficulty level of average and a moderate discrimination level. Students were required to be able to understand that in order to show the shape, center and spread of a quantitative distribution, a graph like a histogram is needed.

Table 7: Graphical Representation

Graphical Representation				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
6	17.49	Average	Moderate	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.
33	29.11	Low Average	Moderate	Understanding that a distribution with the median larger than mean is most likely skewed to the left.
35	48.7	Average	Moderate	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.
2	52.16	Average	Moderate	Ability to recognize two different graphical representations of the same data (boxplot and histogram).
14	58.75	High Average	Strong	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.
15	59.6	High Average	Moderate	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.
34	60.6	Average	Strong	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.
4	66.97	High Average	Strong	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants).
1	78.79	Easy	None	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.
5	79.72	Easy	Strong	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)
3	83.94	Easy	Moderate	Ability to visualize and match a histogram to a description of a variable (neg. skewed distribution for scores on an easy quiz).
12	91.23	Easy	Moderate	Ability to compare groups by comparing differences in averages.
11	94.42	Easy	None	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.

The forth concept analyzed was boxplots, shown in Table 8. The range for percent correct is 52.6%. The item that students had the most difficulty with is item 9 with a corresponding percent correct of 17.0%. The difficulty associated with question 9 is average, and it is moderately discriminatory. Item 9 asked students to identify 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. Thus, the correct response is that it is impossible to determine. Item 8 showed the highest percentage and this corresponds to the highest item for descriptive statistics (see Table 5).

Table 8: Boxplots

Boxplots				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
9	16.97	Average	Moderate	Understanding that boxplots do not provide estimates for percentages of data above or below values except for the quartiles.
10	20.22	Average	Strong	Understanding of the interpretation of a median in the context of boxplots.
2	52.16	Average	Moderate	Ability to recognize two different graphical representations of the same data (boxplot and histogram).
8	69.58	High Average	Moderate	Ability to determine which of two boxplots represents a larger standard deviation.

The fifth concept investigated was data collection and design (see Table 9 below). The range of percent correct is 69.0%. The lowest percentage of correct responses is item 7 with 4.4%. Item 7 is the lowest scoring item on the entire CAOS exam for STAT 2000 students. Item 7 is the hardest question for data collection and design and has no discrimination present between high scorers and low scorers. The concept that is being assessed in item 7 is randomization within an experiment. Students are asked to give the best explanation for the purpose of randomization within an experiment. Item 13 has the highest percentage of correct responses (73.4%). Item 13 is an easier than average question with strong discrimination. It requires students to understand that it is not necessary to have equal sample sizes in order to make comparisons between two groups.

Table 9: Data Collection and Design

Data Collection and Design				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
7	4.41	Hard	None	Understanding of the purpose of randomization in an experiment.
38	31.72	Low Average	Strong	Understanding of the factors that allow a sample of data to be generalized to the population.
22	47.62	Average	Strong	Understanding that correlation does not imply causation.
13	73.42	High Average	Strong	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.

The sixth concept analyzed was probability, see Table 10 below. The range for probability for percent correct is 48.5%. Probability corresponds to the least number of items, with only three items. The item with the lowest percentage of correct response is item number 37. Item 37 is considered a hard question that moderately discriminates between high and low scorers. Item 37 asked students to provide an accurate estimate of the probability of anyone getting at least four out of six tries right by chance, where each try has two outcomes. The one item that students seemed to understand the most out of all the probability items was item 17. Item 17 had a correct response rate of 62.1%, which compared to other concepts' highest scoring items, is not very high. Item 17 requires students to understand the probability associated with expected patterns within samples.

Table 10: Probability

Probability				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
37	13.55	Hard	Moderate	Understanding of how to simulate data to find the probability of an observed value.
36	55.05	Average	Strong	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.
17	62.05	Average	Strong	Understanding of expected patterns in sampling variability.

The seventh concept examined was sampling variability (see Table 11, below). The range for percent correct is 74.4%. Sampling variability had easy, average, and hard questions and also had different discrimination levels. The item with the lowest total correct response percentage is item 32 with a percentage of 9.37. Item 32 does not have any discrimination between low and high scorers and has a difficulty level of hard, corresponding to the hardest item within sampling variability. Item 32 required students to recognize that an estimate of sampling error was needed to conduct an informal inference about a sample mean. The item with the highest percent of correct response is item 18. Item 18 is an easy item with moderate discrimination. Item 18 describes a situation of repeated measurements in which sampling variability must be taken into consideration to pick the correct statement with the smallest amount of variability.

Table 11: Sampling Variability

Sampling Variability				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
32	9.37	Hard	None	Understanding of how sampling error is used to make an informal inference about a sample mean.
16	40.33	Average	Very Strong	Understanding that statistics from small samples vary more than statistics from large samples.
35	48.7	Average	Moderate	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.
34	60.6	Average	Strong	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.
17	62.05	Average	Strong	Understanding of expected patterns in sampling variability.
18	83.76	Easy	Moderate	Understanding of the meaning of variability in the context of repeated measurements and in a context where small variability is desired.

The eighth concept investigated was tests of significance (Table 12). The range for percent of correct responses is 22.8%, which is the lowest range for any of the concepts. Tests of significance only had items of average to high average difficult with discriminations of level mostly moderate with one strong discrimination for item 40. Item 25 had the lowest percentage of correct responses with 48.34 percent. Item 25 was of average difficulty and had a moderate discrimination level. The measured outcome for item 25 was the ability to recognize a correct interpretation of a p-value. Item 24 had the highest percent of correct responses (72.14). It is an easier than average question with moderate discrimination. It required students to understand what conclusions can be made from a statistically significant difference within samples.

Table 12: Tests of Significance

Tests of Significance				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
25	48.34	Average	Moderate	Ability to recognize a correct interpretation of a p-value.
27	50.79	Average	Moderate	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is effective).
23	59.2	High Average	Moderate	Understanding that no statistical significance does not guarantee that there is no effect.
40	61.17	Average	Strong	Understanding of the logic of a significance test when the null hypothesis is rejected.
26	61.45	Average	Moderate	Ability to recognize an incorrect interpretation of a p-value (probability that a treatment is not effective).
19	71.8	High Average	Moderate	Understanding that low p-values are desirable in research studies.
24	72.14	High Average	Moderate	Understanding that an experimental design with random assignment supports causal inference

The last concept examined was confidence levels (Table 13). The items for confidence levels tended to be on the easier scale of difficultly and also had moderate to no discrimination. The lowest percent of correct responses was item 30 with 33.0 percent. Item 30 was a harder than average item (the hardest question for confidence intervals) with no discrimination between high and low scorers. Item 30 required the ability to be able to identify the misinterpretation of a confidence level. Item 31 is the easiest question in the group and also had the highest correct response (83.65). It required students to be able to correctly interpret a confidence interval. Apparently, students were able to identify the correct interpretation of a confidence interval (item 31) but had difficulty being able to identify the wrong interpretation (item 30).

Table 13: Confidence Intervals

Confidence Interval				
Item #	% Correct	Difficulty	Discrimination	CAOS Measured Outcomes
30	32.97	Low Average	None	Ability to detect a misinterpretation of a confidence level (percentage of all possible sample means between confidence limits)
28	41.64	Average	Moderate	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits)
29	59.55	Average	Moderate	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).
31	83.65	Easy	None	Ability to correctly interpret a confidence interval.

5.2 Comparison to national CAOS study

delMas, et al. (2007) report an average percentage correct on the CAOS given as a pretest to a national sample of 763 students as 44.9 percent. The posttest increased 9 percentage points to an average percentage correct of 54.0 percent. The UGA STAT 2000 students showed a similar score with an average percentage correct of 54.34 percent (see Table 4). More recent data from the 17,655 students who have taken the CAOS as a post-test through the CAOS web interface are given in Table 14 (Bob delMas, personal communication). In order to determine if there was a significant difference in the mean scores of the national study and UGA, a one-sample z- test was performed. The large national sample, which included United States undergraduate students enrolled in a 2-year college, 4-year college, or university between 2005 and 2012 who completed a non-calculus-based intro statistics course and answered all 40 questions on the CAOS posttest, was considered to be the base population. The sample of 1202 UGA students was considered to be a sample of all past and future STAT 2000 students at UGA.

Table 14: One-Sample Z-Test

Group	N	Mean	Standard Deviation
National	17655	52.98%	14.44%
UGA	1202	54.35%	10.20%

A one- sample z -test tests the following hypothesis:

H_o: μ_{uga} =52.98

H_a: $\mu_{uga} \neq$ 52.98

The following formula was used to produce the test statistic:

$$z = \frac{\bar{x} - \Delta}{\sigma / \sqrt{n}}$$

where $\bar{x}$ corresponds to the mean of UGA study and Δ corresponds to the mean of the national study, which is taken to represent the population of all undergraduate students in the U.S.. Plugging in the values results in the following equation:

$$z = \frac{54.35 - 52.98}{14.44 / \sqrt{1202}}$$

This results in a test statistic of 3.289. The area of the standard normal curve corresponding to a ***z***-score of 3.289 is 0.000503. This test is two-tailed so the area is doubled and results in a probability of 0.001006. Thus, there is statically significant evidence to reject the null hypothesis and claim the UGA average scores differ significantly from the national average scores.

The difference between the CAOS administration in the UGA STAT 2000 class and the national study reported in delMas, et al. (2007) is the national scores were recorded for both a pretest and a posttest. Due to the fact that UGA STAT 2000 students only took the posttest, there is no method to compare the scores of a pre- and posttest. Although there are no pretest scores for UGA students, the national posttest scores can still be compared to the UGA STAT 2000 students. The five items with the lowest and highest total percentage correct for UGA were compared to the national study (see Table 15 and 16 below).

Table 15: UGA Lowest Percentage vs. National Posttest Scores

Item	National percent correct (rank)	UGA percent correct (rank)	Statistical Concept
7	12.30 (40)	4.40 (40)	Data Collection and Design
39	24.50 (37)	8.57 (39)	Bivariate
32	17.10 (39)	9.32 (38)	Sampling Variability
37	19.50 (38)	13.48 (37)	Probability
9	26.60 (35)	16.89 (36)	Boxplots
6	25.2 (36)	17.49 (35)	Graphical Representation

Table 16: UGA Highest Percentage vs. National Posttest Scores

Item	National percent correct (rank)	UGA percent correct (rank)	Statistical Concept
3	73.20 (9)	83.94 (4)	Graphical Representation
12	85.80 (3)	90.85 (3)	Graphical Representation
11	88.20 (2)	94.34 (2)	Graphical Representation
18	80.60 (5)	83.69 (5)	Sampling Variability
20	92.50 (1)	96.26 (1)	Bivariate
21	83.70 (4)	82.08 (7)	Bivariate

It would appear that the students in the national sample did better the UGA students on all of the lowest scoring items. The item that showed the largest difference in national and UGA is item 39, with a difference of 15.9%. On the items that UGA scored the highest, the students in the national study did not score as high. The item with the largest difference between the national and UGA studies is item 3, with a difference of 10.7%. Although there are differences in average percentage correct between the national and UGA studies, Table 16 and 17 confirm that there are similarities as well. The five lowest percent items for the national posttest were (in order of lowest to greatest percent): 7, 32, 37, 39, and 6. Of those items, 7, 32, 37, and 39 were all lowest scoring items for UGA as well. The five highest percent items for the national posttest were (in order of lowest to greatest percent): 18, 21, 12, 11, and 20. Of those items, 18, 12, 11, and 20 were all highest scoring items for UGA.

In order to analysis internal consistency, the national CAOS determined the Cronbach's alpha. The national CAOS test produced a Cronbach's alpha coefficient of 0.82. Table 17 displays the SAS output for the Cronbach's alpha coefficient.

Table 17: SAS Output for Cronbach Coefficient Alpha

Variables	Alpha
Raw	0.569
Standardized	0.573

There are different standards for an acceptable level of reliability but most suggest having a lower level of 0.5 to 0.7 (Pedhazur and Schmelkin, 1991). Thus, the CAOS test given to the UGA students was judged to have an acceptable internal consistency. The reliability coefficient for the UGA results were, however, lower than those reported by the CAOS research team for the national administration of the CAOS 4. In order to compare the two reliability coefficients for UGA and national, Feldt's test of the equality of two reliability coefficients was performed. The hypothesis test to compare the two reliability coefficients is the following:

H_o: ρ1 =ρ2

H_a: ρ1 <ρ2

Set ρ1 = UGA and ρ2 = National

The formula for the test statistic is the following:

$$W = \frac{1 - \rho 2}{1 - \rho 1}$$

To obtain the test statistic, the reliability coefficients for both the national and UGA study are plugged into the equation:

$$\text{W}= \frac{1-.82}{1-.573} = .42154$$

W follows an *F*-distribution with degrees of freedom ν_1 and ν_2. In order to calculate these degrees of freedom F_1 (UGA) has degrees of freedom: $df_1= (n_1 -1)$ and $df_2= (n_1-1) (k_1-1)$ where n is equal to the number of students and k is equal to the number of items on the assessment. F_2 (national) has degrees of freedom: $df_3= (n_2-1) (k_2-1)$ and $df_4= (n_2-1)$. Feldt used moments of the *F*-distribution to find the following:

$$\text{A}= \frac{df_4}{df_4-2} * \frac{df_2}{df_2-2}$$

$$\text{B}=\frac{(df_1+2)\, df_4^2}{(df_4-2)(df_4-4)df_1} * \frac{(df_3+2)df_2^2}{(df_2-2)(df_2-4)df_3}$$

The resultant degrees of freedom for W are:

$$\nu_1 = \frac{2A^2}{2B-AB-A^2}$$

$$\nu_z = \frac{2A}{A-1}$$

Plugging in the information about n and k for both the UGA and national studies gives the following information for each of the degrees of freedom (Table 18).

Table 18: Degrees of Freedom

df_1	df_2	df_3	df_4
1201	46839	29718	762

Using the information in Table 18 the degrees of freedom were obtained as the following:

$\nu_1 = 1155 \; \nu_2 = 750$

The test statistic W= .42154 with degrees of freedom (1115, 750) produces a p-value of <.0001. Therefore there is evidence to reject the null hypothesis and claim that UGA's reliability coefficient is less than the national's reliability coefficient. Although the UGA's reliability is significantly lower than the national study it should be noted that due to restricted range, reliability will be lower (Allen & Yen, 2001). UGA has a restricted range as the sample includes only UGA STAT 2000 students. On the other hand, the national study had students from many different universities. Therefore, it is not unusual the reliability would be lower for UGA. In addition, UGA's standard deviation for percent answered correct is 10.20% and the national study had a standard deviation of 14.44% (see Table 14). The formula for Cronbach's α is the following:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma^2 \; (Y_i)}{\sigma^2 X}\right)$$

where the n represents the total number of items, $\sum_{i=1}^{n} \sigma^2 \; (Y_i)$ corresponds to the variance for each item, and $\sigma^2 X$ represents the variance of the observed total test scores. Using the formula for Cronbach's alpha it can be concluded the larger the variance the larger the reliability coefficient will be. Thus, the national study will have a larger reliability coefficient based upon the fact the standard deviation was 14.445 compared to UGA's standard deviation of 10.205.

CHAPTER 6

DISCUSSION

In this paper the importance of assessment in statistics education was discussed and illustrated. Assessments provide teachers insight into what topics students understand. Specifically, this paper examined the CAOS assessment. The CAOS assessment is used to provide valuable information on what students appear to learn after one introductory college-level statistics course. The CAOS assessment was broken down into categories of concepts taught in introductory statistics and the common misconceptions associated with those concepts. The nine concepts include: descriptive statistics, bivariate, graphical representation, boxplots, data collection and design, probability, sampling variability, tests of significance, and confidence intervals.

The results of UGA STAT 2000 students' scores indicate that STAT 2000 students are performing at a similar level as comparable students nationwide on the outcomes tested by the CAOS. Most of the higher scoring items for UGA are items that address some type of graphical representation. Items 1, 3, 5, 11,12, 20, and 21 all present a type of graph (boxplot, scatterplot, or histogram) and then ask a corresponding question about interpreting the graph (see Table 19 below). It appears that students at UGA are capable of describing and identifying patterns displayed by the graphs. In addition, students are able to understand the graph's scale and measurement units. These two behaviors correspond to the first two levels of Curcio's three levels of graphical understanding discussed in Section 2.3: reading the graph and reading within the graph. The third and highest level, reading beyond the graph, corresponded to items students did not understand as well. For instance, item 6 required students to read beyond the graph and had a total correct percentage of 17.49. Item 6 required students to know that a histogram is needed to show shape, center, and variability of a distribution of quantitative data. Unfortunately, many students chose the bar graph that was bell-shaped even though, as a bar chart, it cannot be used directly to determine the shape, center, and variability. A person who can read beyond the graph would possess the ability to know what type of graph is needed to determine certain statistics.

Table 19: STAT 2000 Graphical Representation

Item	Rank	% Correct	CAOS Measured Outcomes
20	1	96.42	Ability to match a scatterplot to a verbal description of a bivariate relationship.
11	2	94.42	Ability to compare groups by considering where most of the data are, and focusing on distributions as single entities.
12	3	91.23	Ability to compare groups by comparing differences in averages.
3	4	83.94	Ability to visualize and match a histogram to a description of a variable (neg. skewed distribution for scores on an easy quiz).
21	7	82.08	Ability to correctly describe a bivariate relationship shown in a scatterplot when there is an outlier (influential point).
5	8	79.72	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)
1	9	78.79	Ability to describe and interpret the overall distribution of a variable as displayed in a histogram, including referring to the context of the data.

With the aim of improving STAT 2000, it is recommended that students become exposed to materials that promote learning to read beyond the graph. Curcio describes students who read beyond the graph as students who asked questions about the dataset and can project into the future. Thus, it is important to create activities that encourage students to ask questions about where the data came from, how the data was collected, etc. A suggestion is to use datasets that students would find interesting and relatable. Students may be more inclined to become involved in learning about data that is current and interesting to them.

Students did not show a strong understanding of the statistical concept of data collection and design. In particular items 7, 22, and 38 seemed to be the most difficult for the students (see Table 20 below). Item 7 was the lowest scoring item on the entire CAOS test and addressed the topic of randomization within a study. Students were unable to determine why randomization would be important. Although students showed an understanding of identifying a scatterplot based on variable descriptions (item 20), they were not able to answer item 22, which involved interpreting correlation. Unfortunately, many students did not understand the idea that correlation does not imply causation. In addition, item 38 required students to understand the conditions that are necessary to make generalizations from a sample to a population. Therefore,

Table 20: STAT 2000 Data Collection and Design

Item #	Rank	% Correct	CAOS Measured Outcomes
7	40	4.41	Understanding of the purpose of randomization in an experiment.
38	32	31.72	Understanding of the factors that allow a sample of data to be generalized to the population.
22	28	47.62	Understanding that correlation does not imply causation.

in regards to topics within the concept of data collection and design, students are having difficulty with understanding randomization, correlation, and factors needed to make generalizations about a sample to a population.

These three topics are not uncommon for students to have difficulty understanding (see Section 2.5). Currently, STAT 2000 covers the topic of data collection and design in roughly two class periods. It can be concluded from these results that students do not fully understand the concepts involving data collection and design. Therefore, it is recommended to increase the amount of class time for data collection and design, and incorporate this topic whenever possible throughout the course.

Another area UGA students did not appear to understand well is probability. Although the CAOS only addressed probability with three items (17, 36, and 37), the highest percent was item 17 with only 62.05 percent of students responding correctly (Table 21). It would appear students do not fully understand how to calculate a probability from a two-way contingency table. In addition, students did not show an understanding of how to simulate data to find the probability of an outcome. Students' misunderstanding of probability could be caused by informal experiences with probability or the lack of experience with mathematical laws of

Table 21: STAT 2000 Probability

Item #	Rank	% Correct	CAOS Measured Outcomes
37	37	13.55	Understanding of how to simulate data to find the probability of an observed value.
36	23	55.05	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.
17	15	62.05	Understanding of expected patterns in sampling variability.

probability (see Section 2.6). In order to determine what level students understand probability, it is recommended that a pre-test on probability be given prior to the start of STAT 2000. Once the information is collection on how much students know or do not know about probability, teaching methods can be modified to better promote the learning of probability.

The other six concept areas, descriptive statistics, bivariate data, boxplots, sampling variability, significance testing, and confidence intervals, were very spread out in regards to UGA student performance. It can be concluded that these concepts are not completely lost on students nor completely understood by students. Each of the six concepts would need to be examined in more detail by observing the corresponding CAOS items to see what subtopics are understood or not understood.

Based on the discrimination and difficulty analysis, we have chosen items from the CAOS assessment that we recommend the STAT 2000 coordinator incorporate into the course as either homework or test items. The following items would help the instructors discriminate the students who understand the material from the students that are struggling. For 8 of the statistical concepts measured by the CAOS, there was a subset of 2 to 3 items with good discriminatory properties that covered a range of difficulty levels. For the ninth topic, bivariate data, there was no appropriate subset of questions. This suggests that future work is needed to develop questions to discriminate student understanding of bivariate data.

In the area of graphical representation, there are three items that are appropriate to use as a method to discriminate students (see Table 22). Item 35 has moderate discrimination and average difficulty and requires students to understand how to select the correct sampling distribution for a population. Item 4 has strong discrimination and is easier than average for difficulty. Item 4 requires students to have the ability to match a histogram to a description of a variable. The last item is item 5 and it has strong discrimination and is easy in difficulty. Item 5 describes a variable and asks students to pick the matching histogram that best describes the variable.

Table 22: Sample Items for Graphical Representation

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
35	Moderate	Average	Understanding of how to select an appropriate sampling distribution for a particular population and sample size.
4	Strong	High average	Ability to visualize and match a histogram to a description of a variable (bell-shaped distribution for wrist circumferences of newborn female infants)
5	Strong	Easy	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)

The area of confidence interval has two items that can be used as a method to identity the students who understand the material (see Table 23). Both item 29 and 28 have moderate discrimination and average difficulty and require students to detect a misinterpretation of a confidence interval.

Table 23: Sample Items for Confidence Intervals

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
29	Moderate	Average	Ability to detect a misinterpretation of a confidence level (percentage of population data values between confidence limits).
28	Moderate	Average	Ability to detect a misinterpretation of a confidence level (the percentage of sample data between confidence limits)

In regards to the area of tests of significance, there are two sufficient items that will separate the students that understand and the students that do not understand (Table 24). Item 40 has strong discrimination and an average level of difficulty. Item 40 requires the understanding of when a null hypothesis is rejected. Item 19 has moderate discrimination and is easier than average in difficulty and requires the knowledge that small p-vales are desirable in studies.

Table 24: Sample Items for Tests of Significance

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
40	Strong	Average	Understanding of the logic of a significance test when the null hypothesis is rejected
19	Moderate	High average	Understanding that low p-values are desirable in research studies.

Descriptive statistics has two items that are acceptable student discriminators (Table 25). Item 8 has moderate discrimination and an easier than average level of difficulty. It requires the ability to identity a boxplot with the larger standard deviation. Item 14 has strong discrimination and an easier than average level of difficulty. Item 14 is similar to item 8 in that it asks for the student to understand standard deviation. The only difference is that item 14 is comparing histograms instead of boxplots.

Table 25: Sample Items for Descriptive Statistics

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
8	Moderate	High Average	Ability to determine which of two boxplots represents a larger standard deviation.
14	Strong	High Average	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.

Sampling variability has three items that are appropriate to use as student discriminators (Table 26). Item 16 has a very strong discrimination and average difficulty. It requires students to understand that statistics from small samples vary more than large samples. Item 17 has strong discrimination and average difficulty and requires the knowledge of expected patterns in sampling variability. The last item, item 34, has strong discrimination and average difficulty. It requires students to understand the law of large numbers.

Table 26: Sample Items for Sampling Variability

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
16	Very Strong	Average	Understanding that statistics from small samples vary more than statistics from large samples.
17	Strong	Average	Understanding of expected patterns in sampling variability.
34	Strong	Average	Understanding of the law of large numbers for a large sample by selecting an appropriate sample from a population given the sample size.

Boxplots has only one sufficient item mainly because the concept of boxplot can be seen in various other items (Table 27). Item 10 for boxplots has strong discrimination and average difficulty. It asks students to understand the median for boxplots.

Table 27: Sample Item for Boxplots

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
10	Strong	Average	Understanding of the interpretation of a median in the context of boxplots.

Probability has two items that are acceptable to use as tools to discriminate students (Table 28). Item 37 has moderate discrimination and hard difficulty. It requires students to understand how to find the probability of an observed value by simulating data. Item 36 has strong discrimination and average difficulty and asks students to calculate ratios to find conditional probabilities.

Table 28: Sample Items for Probability

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
37	Moderate	Hard	Understanding of how to simulate data to find the probability of an observed value.
36	Strong	Average	Understanding of how to calculate appropriate ratios to find conditional probabilities using a table of data.

The last area of statistical concepts is data collection and design. Data collection and design has three items that can be used to identify the students that know the material (Table 29). Item 38 has strong discrimination and harder than average difficulty. It requires students to understand the concepts that are need to generalize a sample to the population. Item 13 has strong discrimination and easier than average difficulty. For item 13, students must understand that equal sample sizes are not needed to compare two groups. Finally, item 22 has a strong discrimination and average difficulty level. It requires students to understand that correlation does not imply causation.

Table 29: Sample Items for Data Collection and Design

Item #	Discrimination	Difficulty	CAOS Measured Outcomes
38	Strong	Low average	Understanding of the factors that allow a sample of data to be generalized to the population.
13	Strong	High average	Understanding that comparing two groups does not require equal sample sizes in each group, especially if both sets of data are large.
22	Strong	Average	Understanding that correlation does not imply causation.

It is the goal of any assessment to provide data and information for instructors that can then be used to promote changes in teaching methods. In this study we have concluded that students in STAT 2000 at UGA are developing a similar level of statistical understanding, as measured by the CAOS, as that developed by students across the country. Our students are doing well in understanding graphical representations, but the STAT 2000 coordinator might consider focusing on instruction that helps students to read beyond the data. With regard to the areas in which the students did not perform as well, probability and data collection and design, the results of this study provide two different recommendations. For the area of probability, the coordinator is encouraged to administer a pre-test to students on their knowledge of probability and then tailor instruction to account for the incoming understanding and misconceptions exhibited by students on the pre-test. The STAT 2000 coordinator is encourage to expand the time spent on the teaching of data collection and design if he believes these learning outcomes are important to the development of a UGA student's statistical knowledge.

For the content areas of descriptive statistics, bivariate data, boxplots, sampling variability, significance testing, and confidence intervals, this report suggests that further research is needed to reach a better understanding of the extent of knowledge STAT 2000 students are currently developing about the topics. This report does, however, provide suggestions for the STAT 2000 coordinator for assessment items that should discriminate levels of understanding exhibited by students in all topic areas except bivariate data analysis. Ultimately, students should be able to leave a first year introductory course with the ability to think and reason about statistics effectively. Incorporating more emphasize on these specific concepts is a first step to accomplishing that outcome. We hope that the results of this study will help the STAT 2000 coordinator to achieve this goal at UGA.

REFERENCES

Allen, M. J., & Yen, W. M. (2001). Introduction to measurement theory. (1 ed.). Long grove: Waveland Pr Inc.

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.

American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.).Washington, DC: Author.

Bakker, A. (2003). Reasoning about shape as a pattern in variability. In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3).* [CD] Mount Pleasant, MI: Central Michigan University.

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. Utrecht, The Netherlands: CD-ß Press, Center for Sci. and Math. Education.

Bakker, A. & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–168). Dodrecht, The Netherlands: Kluwer.

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about boxplots? In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education (IASE) Roundtable* (pp. 163–173). Voorburg, The Netherlands: International Statistical Institute.

Batanero, C., Estepa, A. Godino, J. D., & Green, D. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, *27*(2), 151–169.

Batanero, C., Godino, J. D., & Estepa, A. (1998). Building the meaning of statistical association through data analysis activities. In A. Olivier & K. Newstead (Eds.): *Proceedings of the 22nd Conference of the International Group for the Psychology of Mathematics Education* (v. 1, pp. 221-236). University of Stellenbosch, South Africa.

Batanero, C. (2005). Statistics education as a field for research and practice. In *Proceedings of the 10th international commission for mathematical instruction*. Copenhagen, Denmark: International Commission for Mathematical Instruction.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389–396.

Bingham, R. (2001) *Assessment Criteria – A Guide*, Learning and Teaching Institute, Sheffield. Hallam University.

Blair, R. C., Higgins, J. J., Topping, M. E. H., & Mortimer, A. L. (1983). An investigation of the robustness of the t test to unit of analysis violations. *Educational and Psychological Measurement, 43*, 69-80.

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Dordrecht, The Netherlands: Kluwer.

Castro Sotos, A., Vanhoof, S., Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2,* 98-113.

Calhoun, A. W., Guyatt, G. H., Cabana, M. D., Lu, D., Turner, D. A., Valentine, S., & Randolph, A. G. (2008), "Addressing the unit of analysis in medical care studies: A systematic review," *Medical Care, 46* (6), 635-643.

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 199–311.

Curcio, F. R. (1989). *Developing graph comprehension*. Reston, VA: National Council of Teachers of Mathematics.

Davies, N. & Marriott, J. (2010). Assessment and feedback in statistics. In P. Bidgood, N. Hunt, and F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 4–7). Chichester, West Sussex, England: John Wiley & Sons Ltd.

delMas, R. C., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55–82.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Estepa, A., & Sánchez-Cobo, F. (2001). Empirical research on the understanding of association and implications for the training of researchers. In C. Batanero (Ed.), *Training researchers in the use of statistics* (pp. 37-51). Granada, Spain: International Statistical Institute.

Everitt, B. (1996). *Making sense of statistics in psychology: A second-level course*. Oxford Press.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83–96.

Feldt, L. S. (1969). A test of the hypothesis that cronbach's alpha or kuder-richardson coefficient twenty is the same for two tests. *Psychometrika*, 34(3):363–373.

Fidler, F. (2006). Should psychology abandon *p*-values and teach CIs instead? Evidence-based reforms in statistics education. In *Proceedings of the seventh international conference on teaching statistics*. International Association for Statistical Education. Salvador, Brazil.

Friel, S., Curcio, F., & Bright, G. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124-159.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield, (Eds.), *The assessment challenge in statistics education* (pp. 1-14). Voorburg, The Netherlands: International Statistical Institute and IOS Press.

Garfield, J., & Ahlgren, A. (1988). Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education, 19,* 44-63.

Garfield, J., delMas, R., & Chance, B. (1999) *Developing statistical reasoning about sampling distribution.* Presented at the First International Research Forum on Statistical Reasoning, Thinking, and Literacy. BeÕeri, Israel.

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

Garfield, J. (2002) The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10* (3).

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83–92.

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.

Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Halpern, D. (1996), *Thought and Knowledge* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* London: Lawrence Erlbaum Associates.

Hinton, P. (1995). *Statistics explained: A guide for social science students*. Routledge Press.

Hirsch, & O'Donnel, (2001). Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education, 9*(2).

Hoerl, R., & Snee, R. D. (2001). *Statistical thinking, improving business performance*. Duxbury.

Huck, S.,& Cormier, W. (1996). *Reading statistics and research (2nd ed.).* New York, NY: HarperCollins Publishers.
Innabi, H. (1999). Students' judgment of the validity of societal statistical generalization. In A. Rogerson (Ed.), *Proceedings of the international conference on mathematics education into the 21st Century: Societal challenges, issues and approaches.* Cairo.

Kahneman, D., & Tversky, A. (1972), Subjective probability: A judgment of representativeness. *Cognitive Psychology,* 5, 430-454.

Kirk, R. E. (2001). Promoting good statistical practices; Some suggestion. *Educational and Psychological Measurement, 61(2),* 213-218.

Lane-Getaz, S. J. (2008). Introductory and intermediate students' understanding and misunderstanding of *p*-values and statistical significance. *Proceedings of the 11th International Congress on Mathematical Education* (ICMe).

MacGillivray, H. (2004). Coherent and purposeful development in statistics across the education spectrum. In G. Burrill, & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable.* Voorburg, The Netherlands: International Statistical Institute.

MacGillivray, H. (2010). Variety in assessment for learning statistics. In P. Bidgood, N. Hunt, and F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 22). Chichester, West Sussex, England: John Wiley & Sons Ltd.

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.

Morris, E. J. (1997). An investigation of students' conceptions and procedural stills in the statistical topic correlation. *Centre for Information Technology in Education,* Report n. 230. Milton Keynes, U.K: The Open University.

National Council of Teachers of Mathematics (1991), *Professional Standards for Teaching Mathematics*, Reston, VA: Author.

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

Perrett, J. (2012). A case study on teaching the topic "experimental unit" and how it is presented in advanced placement statistics textbooks. *Journal of Statistics Education, 20*(2).

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

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

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L., & Thompson, P. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In. D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465-494). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In Lester, F. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957–1009). Charlotte, NC: Information Age Publishing

Silverman, S. & Solmon, M. (1998). The Unit of Analysis in Field Research: Issues and Approaches to Design and Analysis. *Journal of Teaching Physical Education, 17*, 270-284.

Smolkowski, K. (2012) "Experimental Unit," *Oregon Research Institute website*, http://www.ori.org/~keiths/Files/Tips/Stats_Unit.html.

Starkings, S. (1997). Assessing student projects. (1997). In I. Gal & J. Garfield, (Eds.), *The assessment challenge in statistics education* (pp. 139-151). Voorburg, The Netherlands: International Statistical Institute and IOS Press.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Vallecillos, A., & Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. In L. Puig & A. Gutiérrez (Eds.), *Proceedings of the 20th conference of the International Group for the Psychology of mathematics education* (pp. 271–378). Valencia, Spain:University of Valenciam.

Vallecillos, A., & Batanero, C. (1997). Conceptos activados en el contraste de hipo´tesis estad´ısticas y su comprensio´n por estudiantes universi-tarios [Activated concepts in the statistical hypotheses contrast and their understanding by university students]. *Recherches en Didactique des Mathematiques, 17,* 29–48.

Watson, J.M., & Kelly, B. A. (2004). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics, and Technology Education, 4,* 371-396.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.