A COMPARISON OF PEDAGOGICAL APPROACHES IN INTRODUCTORY STATISTICS

by

BARBARA DOLANSKY

(Under the Direction of Jennifer Kaplan)

ABSTRACT

Over the past 10 years, curricula have been developed that rely on simulation techniques to teach the introductory statistics course. Some statistics educators have argued that the logic of statistical inference should be at the heart of the introductory course, and that using simulation and randomization can lead students to a better understanding of the core ideas of statistical inference. In this study, students' scores on both a revised version of the Comprehensive Assessment of Outcomes in Statistics (CAOS) assessment and on final exam questions will be compared to determine whether students who follow a simulation-based curriculum learn more about statistical inference than students who follow a traditional introductory statistics curriculum.

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

A COMPARISON OF PEDGOGICAL APPROACHES IN INTRODUCTORY STATISTICS

by

BARBARA DOLANSKY

B.A., State University of New York at Albany, 1992

M.A., State University of New York at Albany, 2000

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

2017

A COMPARISON OF PEDAGOGICAL APPROACHES IN INTRODUCTORY STATISTICS

by

BARBARA DOLANSKY

Major Professor:	Jennifer Kaplan
Committee:	Cheolwoo Park
	Lynne Seymour

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

This thesis is dedicated to Robert Seidel. Although it is impossible to list all the ways in which he helped me during this process, it is not an exaggeration to say that this project could not have been completed without his support.

ACKNOWLEDGEMENTS

There are many people I wish to thank. First and foremost, Dr. Jennifer Kaplan, who has been a wonderful mentor and guide throughout this entire process. Her insights, suggestions, and support have been invaluable. I would also like to thank Toni Emery, who encouraged me to pursue this degree and assured me that I could accomplish this goal. I am grateful to Dutchess Community College for granting me sabbatical leave to complete my coursework at UGA, and to the UGA Statistics Department for taking a chance on an older adult student and admitting me to their program. Thank you also to Dr. Cheolwoo Park and Dr. Lynne Seymour for being willing to serve on my thesis committee. I would like to acknowledge Dr. Rob Gould, Dr. Brian Jersky, and Karen Kinard. Attending their CAUSEway workshop in the summer of 2009 is what got me interested in statistics education and started me on this path. Thank you to Dr. Nathan Tintle and his research group for creating the SBI curriculum, inviting me to participate in their research study, and providing data for this project. I would also like to thank my colleagues in the Mathematics and Computer Sciences Department at DCC. They have been completely supportive of my work and are excited to continue statistics education research in the future. Finally, thank you to my family and friends who have always given me unconditional love and support. Without them, none of this would have been possible.

This thesis was completed as part of NSF DUE Grant Number 1323210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Over the past twenty-five years, statistics educators have been advocating for a change in the way the introductory statistics course is taught (Cobb, 1992; Moore, 1997). The Guidelines for Assessment and Instruction in Statistics Education (GAISE) were published in 2005 (Aliaga et al., 2005) and updated in 2016 (GAISE College Report ASA Revision Committee, 2016). Among the six recommendations made in the 2016 report concerning the teaching of introductory statistics are to teach statistical thinking by teaching statistics as an investigative process of problem-solving and decision-making, to focus on conceptual understanding rather than procedural knowledge, to use technology to explore concepts and analyze data, and to use assessments to improve and evaluate student learning. In addition, Cobb (2007) advocated for a change in pedagogy and content in introductory statistics: incorporating the use of randomization and simulation to teach the main ideas of statistical inference. It has been suggested that this approach allows students to grapple with the most difficult ideas in the introductory statistics curriculum throughout the semester so that they will have better success in understanding those key concepts (Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011).

Some work has been done to assess students' understanding of statistical inference in a traditional distribution-based (consensus) curriculum as compared to a randomization-based curriculum. These studies have shown that students learn more about statistical inference with the randomization-based curriculum, and retain the information for a longer period of time (Tintle et al., 2011; Tintle, Topliff, VanderStoep, Holmes, & Swanson, 2012).

For this research project, data were collected to compare students' learning, specifically with regard to statistical inference, in an introductory statistics course when exposed to a simulation-based inference (SBI) curriculum versus the consensus curriculum. The research question that is being addressed with this project is, do students who follow the SBI curriculum learn more about statistical inference than students who follow the consensus curriculum?

CHAPTER 2

LITERATURE REVIEW

Konold and Pollatsek identified the main goal of introductory statistics courses as "to teach ideas central to statistical inference" (2002, p. 260). Even for those who will be consumers, rather than producers, of statistics, the ability to make sense of the results of statistical inference is an important one (Garfield & Ben-Zvi, 2008). The literature has established, however, that many of the ideas necessary to understand inference are difficult for students to grasp. In fact, studies have identified misunderstandings and misconceptions related to statistical inference among teachers and researchers as well. Many areas of misunderstanding related to statistical inference have been identified in the literature, including those involving sampling variability and sampling distributions, those involving confidence intervals, and those involving tests of significance.

2.1 Misconceptions Involving Sampling Variability and Sampling Distributions

In order to understand sampling distributions, students must assimilate several different concepts, including probability, variability, sampling, random variables, and the normal distribution. Although they may understand each concept in isolation, students often have difficulty integrating and applying them to inferential reasoning (Batanero, 2005). In particular, students may confuse the population distribution, sample distribution, and sampling distribution (Castro Sotos, Vanhoof, Noorgate, & Onghena, 2007). If students believe that each individual sample should be representative of the population, they may mistakenly believe that the sampling distribution will look like the population distribution (Chance, delMas, & Garfield, 2004). Some

educators have attempted to alleviate this confusion by presenting students with a visual representation of all three distributions (Saldanha & Thompson, 2002; Garfield & Ben-Zvi, 2008).

Understanding how samples vary is a necessary component of inferential statistics (Garfield & Ben-Zvi, 2008). In a study involving university psychology students, Well, Pollatsek, and Boyce (1990) found that although students seemed to understand that larger samples would represent the population better than smaller samples, they had trouble extending that idea to the relationship between sample size and sampling variability. Chance et al. (2004) noted that students believe sampling distributions for large samples have more variability than sampling distributions for small samples. In a study of over 700 undergraduate students, only about one-third understood that statistics from small samples vary more than statistics from large samples (delMas, Garfield, Ooms, & Chance, 2007). In addition, about half understood expected patterns in sampling variability. Fewer than 20% of students recognized that an estimate of sampling error was needed to conduct an informal inference about a sample mean. Instead, more than half of students based the inference only on the sample standard deviation and did not attend to the sample size and sampling variability.

2.2 Misconceptions Involving Confidence Intervals

Although relatively little research has been done about misconceptions concerning confidence intervals (Castro Sotos et al., 2007), some misconceptions have been identified. While delMas et al. (2007) found that nearly three-quarters of students recognized a correct interpretation of a confidence interval on the Comprehensive Assessment of Outcomes in Statistics (CAOS) posttest, they also saw large percentages of students choose incorrect interpretations of confidence level. For example, more than half of students indicated that the

confidence level is the percentage of all possible sample means between the confidence interval limits. Nearly 57% stated that the percentage of sample data between the confidence interval limits is a valid interpretation of the confidence level. Finally, about one-third of students believed that the confidence level is the percentage of population data values between the confidence interval limits. Thus, delMas et al. (2007) noted that many students who correctly identified the confidence interval interpretation also held misconceptions concerning the meaning of the confidence level. Garfield and Ben-Zvi (2008) identified additional misinterpretations of the confidence level as the chance that the confidence interval includes the sample mean, and the chance that the population mean will be within the confidence interval limits.

Fidler (2006), in a series of experiments with 180 undergraduate psychology and ecology students in Australia, found that 38% described the confidence interval as a set of plausible values for the sample mean, while another 21% indicated they were unsure how to describe the confidence interval. In addition, 73% of students thought that a 90% confidence interval would be wider than a 95% interval (for the same set of data), and nearly half of students chose an incorrect relationship between sample size and confidence interval width. In fact, only 22% gave the correct description of the confidence interval, and only 16% selected the correct relationship between sample size and confidence interval width.

2.3 Misconceptions Involving Hypothesis Tests

There are several common misconceptions regarding interpreting p-values. In general, when conducting a test of significance, the p-value may be correctly interpreted as the probability of seeing results at least as extreme as those actually observed, assuming the null hypothesis is true. Two misinterpretations of p-value are that it is the probability that the null

hypothesis is true, or the probability that the alternative hypothesis is true. In the study by delMas et al. (2007), just over half of students recognized a correct interpretation of p-value at the end of their introductory statistics course, while more than 40% chose an incorrect interpretation of p-value as valid. In particular, 41% of students interpreted a p-value of 0.04 as the probability that the null hypothesis is true, while 47% interpreted the same p-value as the probability that the alternative hypothesis is true.

Additional related misconceptions involve connecting statistical significance to the truth of the hypotheses. Haller and Krauss (2002) gave the results of a hypothetical research study, including a p-value of 0.01, and a set of six statements to 113 psychologists from six universities in Germany. The subjects included professors who taught statistics, professors who did not teach statistics, and psychology students. Each of the six statements was an incorrect interpretation of the study results, and participants were asked to mark whether each statement was true or false. About one-quarter (25.6%) of respondents stated that the null hypothesis had been found to be true, and 43.4% stated that the alternative hypothesis had been found to be true. In addition, 21.2% of the sample thought that the study results showed that the null hypothesis had been disproved, while 15% believed that the alternative hypothesis had been proved. Interpreting the p-value as the probability that the study results occurred by chance or were caused by chance, and the complement of the p-value as the probability that the alternative hypothesis is true, have been identified as prevalent misunderstandings of education researchers when interpreting the results of statistical significance tests (Carver, 1978).

Another area of misunderstanding with respect to hypothesis tests is the interpretation of "reject" and "fail to reject" conclusions. In the study by delMas et al. (2007), when comparing two groups from a randomized experiment, about one-third of students (35.6%) did not

understand that a finding of no statistically significant difference between groups did not guarantee that there was no effect. In addition, only 52% of students chose the correct interpretation when the null hypothesis was rejected in a one-sample test. In particular, more than one-third of students indicated on the posttest that rejecting the null hypothesis means that the null hypothesis is false.

2.4 Addressing Misconceptions

Recognizing the difficulties many students have in understanding the concepts, process, and interpretation of statistical inference, educators have made suggestions concerning ways that statistics instruction can be developed to address misunderstandings and improve students' reasoning and thinking skills. Rossman and Chance (1999) created a list of ten recommendations for teaching statistical inference:

1. Insist on complete presentation and interpretation of results in the context of data.
2. Help students to see the common elements of inference procedures.
3. Always examine visual displays of the data.
4. Always consider issues of data collection.
5. Stress the limited role that inference plays in statistical analysis.
6. Help students to recognize that insignificant results do not necessarily mean that no effect exists.
7. Accompany tests of significance with confidence intervals whenever possible.
8. Present tests of significance in terms of *p*-values rather than rejection regions.
9. Encourage students to use technology to explore properties of inference procedures.
10. Have students perform physical simulations to discover basic ideas of inference.

Batanero (2000) advocated connecting statistical inference to research methodology and experimental design in a general way so that students can see it as part of the overall process of scientific inference. She argued that students should be asked to solve real problems that require them to work through all the phases of the investigation process. Garfield and Ben-Zvi (2008)

presented lessons and classroom activities collected from various statistics educators that cover the main topics in an introductory statistics course. Malone, Gabrosek, Curtiss, and Race (2010) discussed an alternative sequencing to the introductory college-level statistics course (Stat 101) that introduces statistical inference in the third week of the semester and proceeds through the various cases to which statistical inference is applied (e.g., one categorical variable, one numerical variable, two or more categorical variables, etc.). Descriptive statistics, probability, and sampling distributions are discussed within each case so that students are working through these ideas, as well as the logic of statistical inference, multiple times throughout the semester. Malone et al. (2010) conjectured that students would develop a better understanding of the important concepts of statistical inference with this resequencing of topics. Carver (2011) also suggested restructuring Stat 101 to help students increase proficiency with statistical thinking and the logic of inference. In his version, the course is divided into four "mounds": populations and variability, samples and variability, the logic of inferential modeling, and further applications of inference. He characterized course elements that promote statistical thinking and the logic of inference, as well as those that serve client disciplines, as "signal" and everything else as "noise", and encouraged instructors to configure their courses with this in mind.

2.5 Simulation Techniques

One approach to developing students' understanding of statistical inference is through simulation, particularly with respect to sampling distributions. As far back as forty years ago, Simon, Atkinson, and Shevokas (1976) employed the Monte Carlo method with students at three different schools: Stat 101 students at the University of Illinois, General Mathematics students at Polk Community College, and Mathematics for General Education Students at Olivet Nazarene College. Simon et al. (1976) found that, at each school, students taught with randomization

methods outperformed students taught with conventional methods. In addition, students taught using randomization methods had better attitudes towards mathematics and probability and statistics compared to students taught using conventional methods (attitude surveys were only administered at Polk Community College and Olivet Nazarene College).

Konold (1994) reported on his use of software developed specifically to help students better understand probability and statistical inference. Using the software, students create a model and ask the computer to take repeated samples of a specified size in order to determine whether a particular outcome (from a real-life situation) is likely due to random chance. Building the model naturally leads to discussions about whether model assumptions are realistic and what effect this might have on simulated probability estimates, which Konold believed is beneficial to students. Jones, Lipson, and Phillips (1994) claimed that using simulation to help students understand the idea of sampling distributions may not be helpful for statistical inference because there is a disconnect when students move from the empirical (simulated) sampling distribution to the theoretical distribution required for a hypothesis test. Their solution is to have students use Computer Intensive Methods, whereby the simulated distribution is used throughout the entire hypothesis test, including generation of a p-value. Simon (1994) believed simulation and resampling techniques as applied to probability questions are more intuitively understood than their theoretical counterparts, and have the potential to help students (and others) gain insight into otherwise inaccessible problems.

delMas, Garfield, and Chance (1999) conducted an action research project in which students used specially designed software to complete activities intended to improve their understanding of sampling distributions. Through their assessments of students, they found that greater conceptual understanding was achieved when students had to make predictions and then

test them, as opposed to when they were guided through an activity. delMas et al. (1999) concluded that simulation can be a helpful tool for developing statistical thinking, but that students should be required to confront their own misconceptions as part of the process.

In a study of twenty-three part-time adult students, Lipson (2002) used six concept maps to investigate the development of students' conceptions of statistical inference. Students used computer simulations to help them comprehend the notion of sampling distribution, and completed the concept maps over a period of six weeks. Lipson (2002) found that understanding of a particular proposition did not necessarily remain intact from one map to the next. For example, the proposition "The sampling distribution of the sample statistic can be modelled by a known probability distribution" was included by 43% of students on the first map, but only 9% of students on the third map. Like Jones et al. (1994), Lipson (2002) concluded that simulations can aid in students' understanding of sampling distributions but that, without purposeful effort by the instructor, students may not make the necessary connections to the theoretical distributions needed to conduct statistical inference.

Lipson, Kokonis, and Francis (2003) conducted a qualitative study of eight students who participated in a computer simulation activity. They identified four developmental stages (recognition, integration, contradiction, and explanation) through which students must work in order to gain a complete understanding of hypothesis testing. In particular, students had difficulty recognizing the contradiction between the sample result and the claim of the null hypothesis (for a significant result), and explaining that contradiction in terms of its implications for the null hypothesis. Lipson et al. (2003) concluded that simulation activities must be carefully constructed to include questions and features that will help students progress through all four stages of their intellectual development.

2.6 Changes to Introductory Statistics

Over the past twenty-five years, statistics educators have been advocating for a change in the way the algebra-based introductory statistics course (Stat 101) is taught (Cobb, 1992; Moore, 1997). The Guidelines for Assessment and Instruction in Statistics Education (GAISE) were published in 2005 (Aliaga et al., 2005) and updated in 2016 (GAISE College Report ASA Revision Committee, 2016). The 2016 report included the following six recommendations concerning the teaching of introductory statistics at the college level:

1. Teach statistical thinking.
    - Teach statistics as an investigative process of problem-solving and decision-making.
    - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

While the guidelines have taken hold in many Stat 101 courses since 2005, the general outline of the course has largely remained the same (GAISE College Report ASA Revision Committee, 2016). In the traditional curriculum, the material for the course is divided into three sections: descriptive statistics and study design, probability and sampling distributions, and statistical inference (Malone et al., 2010). Simulation, as described in the previous section, has been used in isolated ways to target misunderstandings in particular areas, such as sampling distributions.

Cobb (2007) advocated for a further, and more radical, change to Stat 101. Rather than basing the course on the normal distribution, Cobb contended that Stat 101 should have the logic of inference at its center. In addition, recognizing that computers have obviated the need for much of the distribution-based theory that has been the mainstay of Stat 101 for decades, Cobb

challenged statistics educators to motivate the entire process of statistical inference through simulation and randomization. Holcomb, Chance, Rossman, Tietjen, and Cobb (2010) reported on a randomization-based curriculum in which students are exposed to the entire statistical investigation process throughout the semester, and inference techniques are based on simulations rather than theoretical distributions (see Rossman, Chance, Cobb, & Holcomb, 2008). Holcomb et al. (2010) described classroom activities they developed, as well as classroom experiments designed to help them further refine their curriculum materials. Tintle et al. (2011) also took up the challenge posed by Cobb (2007) and developed a randomization-based introductory statistics curriculum that incorporated the GAISE guidelines. This course introduces statistical inference via randomization tests in week 1 and uses an active learning approach (Tintle, Swanson, & VanderStoep, 2009). Traditional topics that are deemphasized include descriptive statistics and the use of symbolic notation and mathematical equations, although students are required to write about statistical and mathematical ideas in their own words. Probability and sampling distributions are not taught separately, as in the traditional curriculum, but are implicitly covered through instruction on randomization tests. Garfield, delMas, and Zieffler (2012), inspired by Cobb (2007), developed the Change Agents for Teaching and Learning Statistics (CATALST) curriculum, which is based on modeling and simulation (see CATALST, 2012). Students start creating models and repeatedly sampling from their models on the first day of class. They make informal inferences first, before moving on to formal statistical inference using simulation-based methods. Budgett, Pfannkuch, Regan, and Wild (2013), also motivated by Cobb (2007), created new learning trajectories and dynamic visualizations for the randomization approach to teaching hypothesis testing. Their materials, tested in a pilot study with ten students, consist of five teaching activities, including a tactile and computer-based simulation. This learning module was

intended for high school seniors, since Stat 101 students would still be required to learn normal-based inference to satisfy client disciplines. Roy et al. (2014) created an updated and combined version of the work of Rossman et al. (2008) and Tintle et al. (2011) with the following key features:

- Use the spiral approach to the statistical investigation process
- Use simulation/randomization-based methods to introduce statistical inference
- Focus on the logic and scope of inference
- Integrate exposition, examples, and active explorations
- Integrate easy-to-use web-based applets for carrying out analyses
- Always use real data from genuine studies that matter

They contend that this new version of Stat 101 has the following advantages over the traditional curriculum: It does not rely on a formal discussion of probability, and hence can be used to introduce statistical inference as early as week 1. It has a lot of opportunity for activity/exploration-based learning and allows one to revisit the entire statistical investigation process over and over again in new contexts. Finally, interpreting and evaluating the p-value becomes much easier when students experience it as a long-run probability through simulation.

2.7 Assessment of Simulation-Based Inference

In her review of the literature concerning using simulation to teach statistics, Mills (2002) commented that there was very little empirical research to support the use of computer simulation methods (CSM). Since then, instructors who have developed simulation-based curricula as described in the previous section have made an effort to assess whether students are experiencing significant learning gains.

The Comprehensive Assessment of Outcomes in Statistics (CAOS) test (delMas, et al., 2007) was developed to measure students' conceptual understanding, as well as to pinpoint

misconceptions, after a course in introductory statistics. In the fall of 2005 and spring of 2006, the test was administered to over 1500 students from 33 post-secondary schools in 21 states. Although a statistically significant increase was found in percentage correct from pretest to posttest, the average percentage correct on the posttest was only 54%. Of the ten posttest questions related to confidence intervals and significance testing, more than 60% of students answered correctly on four of the items and fewer than 60% of students answered correctly on six of the items. Among the authors' conclusions were that students did not demonstrate an understanding of confidence intervals, and that many students hold contradictory interpretations of the p-value. Since its initial administration, almost 14,000 students have taken the CAOS test, and the average (overall) percent correct on the posttest has remained stable (Garfield et al., 2012).

Holcomb, Chance, Rossman, and Cobb (2010) created preliminary assessment materials to evaluate students' understanding of p-value and statistical significance after using randomization-based curriculum materials. Initial results suggest that they are making progress, both in terms of using randomization-based methods and with the assessment instruments they have developed, and they continue to revise classroom activities and assess student understanding. Tintle et al. (2011) used the CAOS test to compare students' understanding of statistical inference in a distribution-based curriculum as compared to a randomization-based curriculum at Hope College. Pretests and posttests were given in eight sections of a traditional course in Fall 2007, and to eight sections of a randomization-based course in Fall 2009. The researchers also compared their results to a national sample (delMas et al., 2007). While the majority of questions showed no difference, significant improvement was found for students in the randomization-based curriculum on questions concerning tests of significance, simulation,

and the purpose of randomization. In a second study (Tintle et al., 2012), students from the Fall 2007 and Fall 2009 cohorts at Hope College took the CAOS test four months after the end of the semester. The results showed that students using the new curriculum demonstrated higher levels of retention, particularly on questions related to tests of significance and data collection/design.

In their pilot study, Budgett et al. (2013) had students complete a pretest, posttest, and post-task after using learning trajectories that included randomization and simulation. They concluded that tactile simulations are an important component that can help students understand the computer simulations. In addition, they believe students need help in moving from the simulated data to an appropriate inferential conclusion. Overall, they found that using randomization made some of the most important underlying concepts of hypothesis testing more accessible to students, and that students learned more about statistical inference than when theoretical methods were used.

Tintle et al. (2014) administered pretests and posttests using CAOS to students at Dordt College to compare outcomes from a traditional curriculum versus a randomization curriculum. Difference in scores from pretest to posttest were twice as high for students in the randomization curriculum. Tintle et al. (2014) then developed a new assessment instrument, which included a mix of CAOS and other multiple-choice questions. Pretests and posttests were administered to students in seventeen sections of statistics, taught by sixteen different instructors at eleven different institutions. Note that all students to whom this instrument was administered were using the preliminary version of Tintle et al. (2016). Significant learning gains were seen overall as well as across six of seven subscales.

Maurer and Lock (2016) conducted a randomized experiment in which students enrolled in introductory statistics courses at Iowa State University were randomly assigned to either a

simulation-based curriculum (see Lock, Lock, Morgan, Lock, & Lock, 2013) or a traditional curriculum. The two curricula were identical until week 9 of the course, after the midterm exam. In the second part of the semester, curricula were similar except that students in the simulation-based curriculum learned the main concepts of inference using randomization methods before also learning theory-based methods. Data collected included student responses to questions from the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) database on the topics of confidence intervals and hypothesis testing (ARTIST, 2006). Although students assigned to the simulation-based curriculum were found to have statistically significantly higher learning outcomes for topics related to confidence intervals, Maurer and Lock (2016) determined that their study was unable to draw reliable conclusions about whether simulation-based methods are more successful for teaching statistical inference.

Following on the work described above, data were collected for this research project to compare students' learning, specifically with regard to statistical inference, in an introductory statistics course when exposed to a simulation-based inference (SBI) curriculum versus a traditional (consensus) curriculum. In particular, this project attempts to determine whether students who follow the SBI curriculum learn more about statistical inference than students who follow the consensus curriculum.

CHAPTER 3

DATA COLLECTION AND METHODS

3.1 Settings and Subjects

Dutchess Community College (DCC) is located in Poughkeepsie, NY, about halfway between New York City and Albany in an area known as the Hudson Valley. DCC is part of the State University of New York (SUNY) system, and enrolls approximately 9500 full-time, part-time, and concurrent enrollment students. Dutchess offers Associate Degrees (A.A., A.S., and A.A.S.) as well as academic certificates in over 60 different areas. One-third of all Dutchess County high school students who attend a college enroll at DCC, and there is on-campus housing that accommodates about 450 students. As of Fall 2015, 61% of DCC students were White, 19% Hispanic, and 14% Black non-Hispanic. In addition, 75% were aged 21 or less, 16% were aged 22 – 29 years, 9% were aged 30 years or older, and 54% were female.

DCC offers one statistics course, MAT118 Elementary Statistics, through its Mathematics and Computer Sciences department. There were 56 sections of MAT118 taught during the 2015–2016 academic year, including 10-week flex course sections, an online section, and day and evening sections at both the main and DCC South campus locations in the fall, spring, and summer semesters. This research project was conducted on three sections of MAT118 during the Spring 2016 semester. Section 1 met on Monday, Wednesday, and Friday from 10–10:50am and had access to a class set of Chromebook laptops. Section 2 met on Monday, Wednesday, and Friday from 11–11:50am in a computer lab classroom. Section 3 met on Tuesday and Thursday from 11:00am–12:15pm and had access to a class set of Chromebook

laptops. Sections 1 and 2 followed the consensus curriculum and used a traditional text (Gould & Ryan, 2016) with StatCrunch (https://www.statcrunch.com/), a web-based data analysis software. Section 3 followed a simulation-based inference (SBI) curriculum and used a text (Tintle et al., 2016) and web-based applets (http://www.rossmanchance.com/ISIapplets.html) specifically designed for that approach.

In Sections 1 and 2, students followed a traditional sequence, with the first two and a half weeks spent on organizing and classifying data, and a graphical understanding of shape, center, and variation. The next two and a half weeks covered measures of center and variation, the empirical rule, and linear regression. Three weeks were then spent on probability and the normal distribution, and the final five weeks covered the Central Limit Theorem for sample proportions, confidence intervals for a single categorical variable, and hypothesis testing for a single categorical variable. The students were given four unit exams during the course of the 15–week semester, as well as a cumulative final exam. There were five homework assignments given, and a participation grade was calculated at the end of each unit. This was based partly on attendance and partly on students' responses to surveys that were given online on an approximately weekly basis.

The curriculum for Section 3 was very different from that of Sections 1 and 2. The first four and a half weeks were spent on a brief overview of the statistical investigation process, summarizing a distribution through shape, center, variability, and outliers, and hypothesis tests for a single categorical variable using both simulation and theory-based methods. The next five weeks covered generalizing from random samples, hypothesis tests for a single quantitative variable using simulation and theory-based methods, type 1 and type 2 errors, and confidence intervals for a single proportion and a single mean using simulation and theory-based methods.

The final three and a half weeks were spent on association and confounding, observational studies versus experiments, and hypothesis tests and confidence intervals for two categorical variables using both simulation and theory-based methods. The students were given three unit exams during the course of the semester, as well as a cumulative final exam. There were seven quizzes given, and a participation grade was calculated at the end of each unit. This was based partly on attendance and partly on students' responses to surveys that were given online on an approximately weekly basis. At the end of the semester, in the 14th week, the students were given four chapter investigations as a way to review for the final exam.

There were 21 students registered in Section 1, 22 students registered in Section 2, and 23 students registered in Section 3. In Section 1, two students could be characterized as "stopped attending", and an additional two students did not take the final exam. In Section 2, two students did not take the final exam. In Section 3, six students could be characterized as "stopped attending", and an additional two students did not take the final exam. Standard IRB protocols were followed at both DCC and UGA, and the number of students who gave consent to use their data was 15 in Section 1, 19 in Section 2, and 12 in Section 3. A total of 45 students had final exam data that could be analyzed, 33 in the consensus cohort and 12 in the SBI group. In addition, 66 students took a pretest, and 66 students also took a posttest (these tests are described in the next section). Of these, 2 were eliminated due to repeated attempts, and another 19 did not give informed consent. Of the 45 remaining, 10 took only one of the two tests but not both, and an additional student had to be removed for missing data. This resulted in a sample of size 34, with 27 students in the consensus curriculum sections and 7 in the SBI section. Demographic information was gathered from the posttest (Table 1).

Table 1: Demographic information based on posttest responses

Gender		**Student Status**	
Female	19	Freshman in college	24
Male	16	Sophomore in college	5
Ethnicity		Other	6
White	23	**Major**	
Black, African American, or Negro	5	Social sciences	20
Hispanic, Latino or Spanish origin	5	Natural and applied sciences	6
American Indian or Alaska Native	2	Arts and humanities	4
Asian	2	Other	5
First–generation college student	20		

3.2 Instruments

Students were given a pretest at the beginning of the semester, and a posttest at the end of the semester. The tests were administered outside of class, online via SurveyMonkey, during the first (pretest) and last (posttest) week of classes. Students were told the test should take about 45 minutes to complete, and that they should do their best but not spend time looking in their notes or on the Internet for help. There was no time limit, so students could conceivably take as long as they wanted to complete the tests. Students were given a 100% homework grade for completing each test regardless of the percentage of questions answered correctly.

The first part of the test consisted of items from the Survey of Attitudes Toward Statistics (SATS–36) (Schau, 2003), which measures six attitude components. A brief description of each component along with an example item follows.

- Affect – students' feelings concerning statistics (6 items)

  Example: "I will feel insecure when I have to do statistics problems"

- Cognitive Competence – students' attitudes about their intellectual knowledge and skills when applied to statistics (6 items)

  Example: "I will find it difficult to understand statistical concepts"
- Value – students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life (9 items)

  Example: "Statistics is not useful to the typical professional"
- Difficulty – students' attitudes about the difficulty of statistics as a subject (7 items)

  Example: "Statistics is a complicated subject"
- Interest – students' level of individual interest in statistics (4 items)

  Example: "I am interested in using statistics"
- Effort – amount of work the student expends to learn statistics (4 items)

  Example: "I plan to work hard in my statistics course"

Responses are along a 7-point Likert scale, from 1: Very Strongly Disagree to 4: Neutral to 7: Very Strongly Agree. Approximately half of the items are worded positively and the other half worded negatively. Verb tense was changed from the pretest to the posttest (e.g., "I will like statistics" versus "I liked statistics").

The second part of the test was a modified version of the CAOS test (delMas et al., 2007) and consisted of 32 multiple choice or multiple select questions. The questions cover six topic areas: Descriptive Statistics (7 items), Data Collection & Design (2 items), Scope of Conclusions (2 items), Sampling Variability/Simulation (6 items), Confidence Intervals (6 items), and Tests of Significance (9 items). In some cases, questions from the CAOS test were modified in context or wording. In other cases, multiple choice CAOS questions concerning interpretation of conclusions were changed to Valid/Invalid options, or vice versa. In addition, seven questions

on this instrument are not on the CAOS test. Most of these address whether students can attend to sample size (small sample size and/or different sample sizes between groups) when interpreting a p-value or interpreting a result that is given as "statistically significant". This test will be referred to as the revised CAOS test, or R–CAOS. All questions referred to in the Results and Discussion chapters, along with their answer choices, can be found in Appendix A.

In addition to the R–CAOS, students in both the consensus and SBI groups were given a comprehensive final exam that included 32 multiple choice questions, 28 of which were the same in both cohorts. Four questions (#23, 25, 28, and 29) were about confidence intervals. These questions addressed the relationship between confidence level and confidence interval width, the relationship between sample size and confidence interval width, the purpose of confidence intervals, and the interpretation of confidence level. Six questions (#7, 12, 16, 20, 30, and 31) were about hypothesis tests. These questions addressed the relationship between p-value and *z*-score, the definition of p-value, whether a p-value can be negative, the fact that hypotheses are statements about parameters, and the interpretation of p-value with and without context. On most questions, common misconceptions were given as distractors. It is worth noting that students were also given similar questions prior to the final exam, on quizzes, practice exams, and unit exams (see Table 2). There were some notable differences in the number of similar questions administered during the semester: Students in the consensus curriculum were not asked about the relationship between a p-value and a z-score, while SBI students were asked about this relationship three times. Students in the SBI curriculum were not asked about the relationship between hypotheses and parameters and were asked to interpret p-values in context fewer times than the students in the consensus curriculum. All questions (final exam questions and similar questions), along with their answer choices, can be found in Appendix A.

Table 2: Number of similar questions given on assessments prior to the final exam

	Curriculum	
Subtopic	Consensus	SBI
Relationship between confidence level and width	3	2
Relationship between sample size and width	3	1
Purpose of confidence intervals	2	2
Interpretation of confidence level	2	1
Relationship between p-value and *z*-score	0	3
Definition of p-value	2	1
Hypotheses are statements about parameters	3	0
Interpretation of p-value (in context)	5	2
Interpretation of p-value (no context)	1	2

3.3 Analysis

In order to answer the research question, answers on the R–CAOS and final exam will be examined. The average number of questions correct on the R-CAOS will be compared for the pretest and the posttest, as well as the difference in average percent correct. A simulation analysis as well as a matched-pairs t-test will be used to determine whether the difference in mean scores is significant for each curriculum group. A simulation analysis as well as a two-sample t-test on the difference of the paired differences in scores will be used to determine whether the difference in means is larger for the SBI group. In addition, the two curriculum groups will be compared on each topic area: sampling variability, confidence intervals, and hypothesis tests on the R-CAOS, and confidence intervals and hypothesis tests on the final exam. Finally, question by question analyses will be conducted within the topic areas on both the R-CAOS and the final exam to compare the SBI group to the consensus cohort. The Matched Pairs and Multiple Means applets from the *Introduction to Statistical Investigations* website (available

at http://www.rossmanchance.com/ISIapplets.html) will be used to conduct simulation analyses, while JMP will be used to conduct parametric analyses.

CHAPTER 4

RESULTS

The purpose of this research project is to determine whether introductory statistics students who follow a curriculum using simulation-based inference (SBI) learn more about statistical inference than students following a more traditional (consensus) curriculum. Specifically, the project focuses on the topics of sampling variability, confidence intervals, and hypothesis tests. This chapter contains the results of student-level data collected using the Revised CAOS (R-CAOS) instrument and analogous questions posed to the two groups of students on the course final exams.

4.1 Analysis of the Revised CAOS

In this section, the results from the R-CAOS instrument are presented. First, overall averages on the pretest and posttest will be compared for the two curriculum groups. Then, the average percent correct will be compared on each of the three topic areas related to statistical inference. Finally, question by question analyses will be conducted within each topic area.

4.1.1 Aggregate Results

There were 32 questions on the R-CAOS, which was administered to students at both the beginning and end of the semester. Posttest means were higher than pretest means for both cohorts (Figure 1, Table 3). A matched-pairs test was conducted separately for each curriculum to determine whether there was a statistically significant increase in average percent correct from pretest to posttest (Table 4). The p-values from both simulation (10,000 repetitions) and the matched-pairs t-test for the consensus cohort indicate that this group scored significantly higher

on the posttest than the pretest. However, the 95% confidence interval of (0.003, 0.154) indicates that the amount of increase may not be practically important. The data from the SBI cohort provided only moderate evidence of a statistically significant increase from pretest to posttest. Since the sample size was so small, the p-value from the simulation method may be more reliable than the p-value from the matched-pairs t-test.

In addition, a two-sample test using simulation with 10,000 repetitions (p-value = 0.335), as well as a two-sample t-test (p-value = 0.325), found no significant evidence that students in the SBI curriculum had a greater difference in average percent correct than students in the consensus cohort.

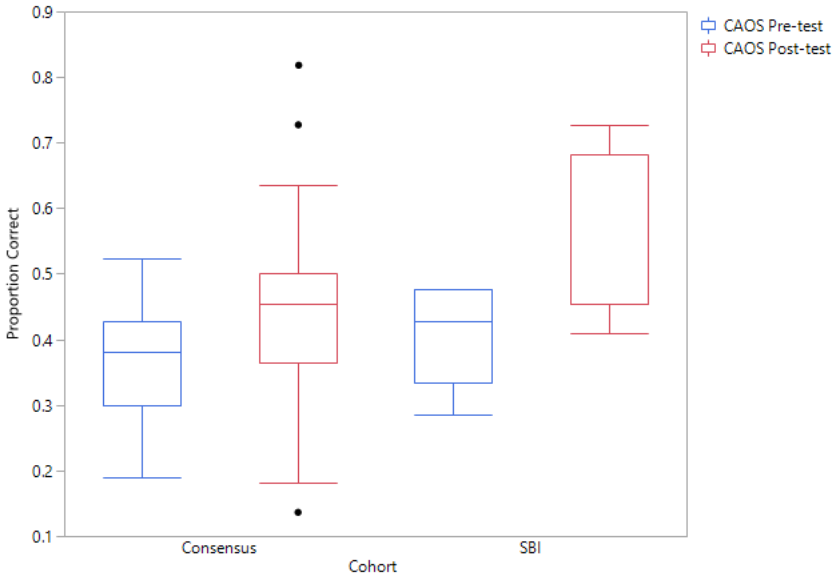

Figure 1: Comparison of R-CAOS pretest and posttest scores

Table 3: Aggregate comparison of R-CAOS scores

		Average percent correct		Difference in average percent correct
	n	Pretest Mean (SD)	Posttest Mean (SD)	Mean (SD)
Consensus	27	36.4 (9.8)	44.1 (14.9)	7.7 (18.8)
SBI	7	41.5 (7.6)	52.6 (12.5)	11.1 (16.7)

Table 4: p-values for difference in average percent correct on R-CAOS

	n	Simulation	Matched-pairs t-test
Consensus	27	0.022	0.022
SBI	7	0.076	0.065

4.1.2 R-CAOS Comparison by Topic

On the three inference topic areas (sampling variability, confidence intervals, and hypothesis tests), neither curriculum group did particularly well on the R-CAOS posttest (Figure 2, Table 5). Recall from Section 3.2 that the test was taken outside of class and a 100% homework grade was given simply for completing the test. It is unknown what effect the unmonitored environment and low stakes value of the test might have had on student effort. However, scores on the confidence interval and hypothesis test topics on the final exam were not markedly different (see Table 12).

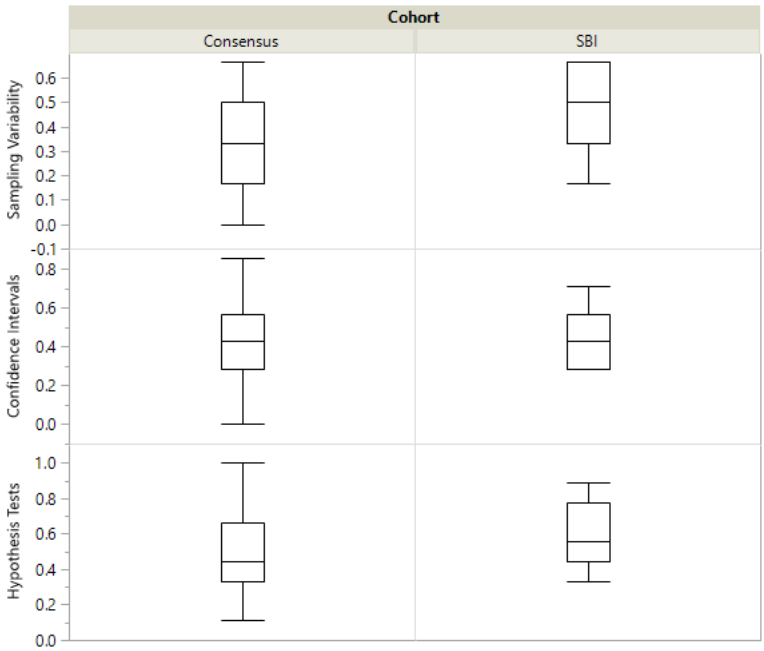

Figure 2: R-CAOS posttest average proportion correct

Table 5: R-CAOS posttest average percent correct by topic

	n	Sampling Variability (SD)	Confidence Intervals (SD)	Hypothesis Tests (SD)
Consensus	27	35.3 (21.2)	43.6 (19.3)	50.4 (24.2)
SBI	7	50.0 (19.2)	44.9 (17.4)	60.3 (19.1)

For each topic area, the difference in average percent correct from pretest to posttest was calculated for each curriculum. This difference was higher for students in the SBI curriculum for both the sampling variability and tests of significance topics, while the difference was higher for the consensus cohort for the confidence intervals topic (Table 6). In order to determine whether any of the differences were statistically significant, a simulation test (10,000 repetitions) and a two-sample t-test were conducted for each topic. Note that the p-values for the sampling variability topic are noticeably different for the two tests (Table 7). This was the only case

where the sample data (consensus curriculum) did not pass the Shapiro-Wilk Goodness-of-Fit test for normality (p-value = 0.0132). Thus, the p-value from the simulation analysis is more reliable than the p-value from the two-sample t-test. These results indicate that there is not statistically significant evidence that students in the SBI cohort had a greater difference in average percent correct than students following the consensus curriculum for any of the three topic areas.

Table 6: Comparison of R-CAOS scores by topic

		Difference in average percent correct		
	n	Sampling Variability (SD)	Confidence Intervals (SD)	Hypothesis Tests (SD)
Consensus	27	0.7 (29.3)	12.7 (26.5)	6.4 (25.3)
SBI	7	19.0 (20.2)	1.2 (23.0)	14.3 (23.8)

Table 7: p-values for difference in average percent correct on R-CAOS by topic

	Sampling Variability	Confidence Intervals	Hypothesis Tests
Simulation	0.0732	0.8585	0.2189
t-test	0.0379	0.8602	0.2291

4.1.3 R-CAOS Question by Question Comparison for Sampling Variability

Two questions (#34 and #44) on the R-CAOS assess students' understanding of the relationship between sample size and sampling variability. Six out of seven students (85.7%) in the SBI cohort answered question 34 incorrectly on the posttest, and only one of these students (16.7%) had answered correctly on the pretest. One student out of seven (14.3%) answered the question correctly on the posttest after answering incorrectly on the pretest. In the consensus cohort, more than three-quarters (77.8%) of students answered question 34 incorrectly on the posttest, and 14.3% of these had answered correctly on the pretest. Three students (11.1%)

answered the question correctly on the posttest after answering incorrectly on the pretest. A smaller percentage of students in each cohort selected the misconception noted by Chance et al. (2004), that sampling distributions for large samples have more variability than sampling distributions for small samples, on the posttest than the pretest (Table 8). It is interesting to see that none of the SBI students chose this particular misconception on the posttest, compared to 42.9% on the pretest. In the consensus curriculum, only one fewer student chose this misconception on the posttest as compared to the pretest.

Table 8: Misconception: larger samples produce more sampling variability

Percentage of students with the misconception			
Consensus (n = 27)		SBI (n = 7)	
Pretest	Posttest	Pretest	Posttest
18.5	14.8	42.9	0.0

Question 44 does not ask about sampling variability directly, but assesses whether students understand that a proportion from a larger sample is more unusual than the same proportion from a smaller sample. In the SBI group, only two students (28.6%) answered this correctly on the posttest, both of whom had answered incorrectly on the pretest. In addition, one student answered incorrectly on the posttest after giving the correct answer on the pretest. In the consensus cohort, over half of students (55.6%) chose the correct answer on the posttest, and 60% of these had chosen an incorrect answer on the pretest. Another 11.1% of students moved from a correct answer on the pretest to an incorrect answer on the posttest. The most common incorrect answer revealed students' misconception that the same proportion from different sample sizes are equally convincing evidence of an effect. In the SBI cohort, one fewer student

chose this answer on the posttest than the pretest, compared to one additional student from the consensus curriculum (Table 9).

Table 9: Misconception: the same proportion from different samples sizes are equally likely

Percentage of students with the misconception			
Consensus (n = 27)		SBI (n = 7)	
Pretest	Posttest	Pretest	Posttest
37.0	40.7	57.1	42.9

4.1.4 R-CAOS Question by Question Comparison for Confidence Intervals

Questions 18 through 20 ask students to determine whether given interpretations of a confidence interval for a mean are valid or invalid. The interpretation in question 19 is valid, while the other two interpretations are not. About one-third (34.6%) of consensus students and almost three-quarters of SBI students (five out of seven) correctly identified the valid interpretation on both the pretest and the posttest (Table 10). An additional 26.9% of the consensus cohort and 14.3% (one student) of the SBI group moved from an incorrect response on the pretest to a correct response on the posttest (Table 10). However, a large proportion of students in both cohorts stated on the posttest that the incorrect interpretations given in questions 18 and 20 were valid, and a fair proportion of these students had correctly chosen the invalid option on the pretest (Table 10). Still, over one-fourth of consensus students (25.9% on question 18 and 30.8% on question 20) recognized the invalid interpretation on the posttest after choosing the valid option on the pretest (Table 10). In the SBI cohort, none of the students moved from an incorrect response to a correct response from pretest to posttest on these two questions (Table 10).

Table 10: Percent correct on confidence interval interpretations
C = Correct response I = Incorrect response

		One randomly selected data value will be within CI limits (Q 18) Posttest			
		Consensus (n = 27)		SBI (n = 6)	
		C	I	C	I
Pretest	C	3.7	14.8	16.7	16.7
	I	25.9	55.6	0.0	66.7

		Correct Interpretation (Q 19) Posttest			
		Consensus (n = 26)		SBI (n = 7)	
		C	I	C	I
Pretest	C	34.6	15.4	71.4	0.0
	I	26.9	23.1	14.3	14.3

		Mean from one random sample will be within CI limits (Q 20) Posttest			
		Consensus (n = 26)		SBI (n = 6)	
		C	I	C	I
Pretest	C	0.0	26.9	33.3	16.7
	I	30.8	42.3	0.0	50.0

Question 46 assesses whether students understand the relationship between sample size and confidence interval width, and confidence level and confidence interval width. In the SBI cohort, five students (71.4%) correctly identified on the posttest that increasing the sample size would produce a narrower confidence interval, and four of these students also answered correctly on the pretest. None of the SBI students gave the correct answer on the pretest but the incorrect

answer on the posttest. In the consensus group, 63.0% understood this concept on the posttest, 64.7% of whom had given the correct response on the pretest. In addition, 14.8% of students answered correctly on the pretest but incorrectly on the posttest. On the posttest, three SBI students (42.9%) understood that increasing the confidence level would not decrease the width of the confidence interval. All of these students answered incorrectly on the pretest. In addition, one student (14.3%) gave the correct answer on the pretest but the incorrect answer on the posttest. In the consensus cohort, 57.7% of students correctly identified the relationship between confidence level and confidence interval width on the posttest, and 80% of these had answered incorrectly on the pretest. Only one consensus student (3.9%) answered this question correctly on the pretest but incorrectly on the posttest.

4.1.5 R-CAOS Question by Question Comparison for Hypothesis Tests

Question 27 assesses whether students understand that a small p-value is needed for a statistically significant result. Six out of seven students in the SBI cohort (85.7%) answered this question correctly on the posttest but incorrectly on the pretest. The student who answered incorrectly on the posttest had answered correctly on the pretest. In the consensus group, 44.4% of students answered this question correctly on the posttest, and 91.7% of these gave an incorrect response on the pretest. Of the 55.6% who answered incorrectly on the posttest, 80% had also chosen the incorrect response on the pretest. In addition, 7.4% of consensus students on the pretest and 18.5% of consensus students on the posttest stated that the magnitude of a p-value has no impact on statistical significance. None of the SBI students chose this option on either the pretest or the posttest.

Questions 28 through 30 ask students to determine whether given interpretations of a p-value for a drug test study are valid or invalid. The interpretation in question 29 is valid, while

the other two interpretations are not. A similar proportion of students (55.5% in the consensus group and 57.2% in the SBI cohort) correctly identified the valid interpretation on the posttest (Table 11). In addition, about half of those in each curriculum who chose the correct option on the posttest had answered incorrectly on the pretest. On question 28 on the posttest, 66.6% of consensus students compared to 28.6% of SBI students incorrectly stated that the p-value is the probability that the alternative hypothesis is true (Table 11). None of the SBI students, compared to 22.2% of consensus students, moved from a correct response on the pretest to an incorrect response on the posttest for this question (Table 11). On question 30 on the posttest, 59.2% of the consensus group and 42.9% of the SBI cohort incorrectly stated that the p-value is the probability that the null hypothesis is true (Table 11). In the consensus group, 43.8% of the students who answered this question correctly on the posttest had given an incorrect answer on the pretest (Table 11). By contrast, all of the SBI students who correctly identified this invalid interpretation on the posttest had also answered correctly on the pretest (Table 11). None of the SBI students, compared to 22.2% of consensus students, moved from a correct response on the pretest to an incorrect response on the posttest for this question (Table 11).

Table 11: Percent correct on p-value interpretations
C = Correct response I = Incorrect response

		Probability H_a is true (Q 28) Posttest			
		Consensus (n = 27)		SBI (n = 6)	
		C	I	C	I
Pretest	C	3.7	22.2	42.9	0.0
	I	29.6	44.4	28.6	28.6

		Correct Interpretation (Q 29) Posttest			
		Consensus (n = 27)		SBI (n = 7)	
		C	I	C	I
Pretest	C	25.9	14.8	28.6	28.6
	I	29.6	29.6	28.6	14.3

		Probability H_0 is true (Q 30) Posttest			
		Consensus (n = 27)		SBI (n = 6)	
		C	I	C	I
Pretest	C	33.3	22.2	42.9	0.0
	I	25.9	18.5	0.0	57.1

Question 43 also asks for an interpretation of a p-value. Four out of seven (57.1%) of SBI students answered this question correctly on the posttest, all of whom had given an incorrect answer on the pretest. One SBI student (14.3%) gave an incorrect answer on the posttest after correctly interpreting the p-value on the pretest. In the consensus cohort, one-third of students (33.3%) correctly interpreted the p-value on the posttest, and about half of these (55.6%) had given an incorrect answer on the pretest. Of the 66.7% of consensus students who answered

incorrectly on the posttest, one-third (33.3%) had identified the correct interpretation on the pretest. In the SBI group, two students on the pretest and two students on the posttest (28.6%) indicated that the small (0.004) p-value meant that the alternative hypothesis had been proven. This same response was selected by 22.2% of consensus students on the pretest and 11.1% of consensus students on the posttest. In addition, three SBI students (42.9%) on the pretest and one SBI student (14.3%) on the posttest indicated that the study result likely happened by chance and supported the null hypothesis. This response was chosen by 11.1% of consensus students on the pretest and 33.3% of consensus students on the posttest. Finally, 29.6% of consensus students on the pretest and 22.2% of consensus students on the posttest believed that the small sample size (n = 15) meant that there was no conclusive evidence for either hypothesis. This response was selected by one SBI student (14.3%) on the pretest and no SBI students on the posttest.

4.2 Analysis of Final Exam Questions

In this section, results from the final exam questions involving confidence intervals and hypothesis tests are presented. First, the cohorts will be compared on each of the topic areas. Then, question by question analyses will be conducted within each topic area.

4.2.1 Final Exam Comparison by Topic

For each topic area, the average percent correct on the final exam in each curriculum was calculated. The average percent correct for both confidence intervals and hypothesis tests was higher for the students in the consensus curriculum (Figure 3), although the 95% confidence intervals overlap in each case (Table 12). It is interesting to note that the only group that had an average percent correct above 50% was the consensus group for the hypothesis tests topic.

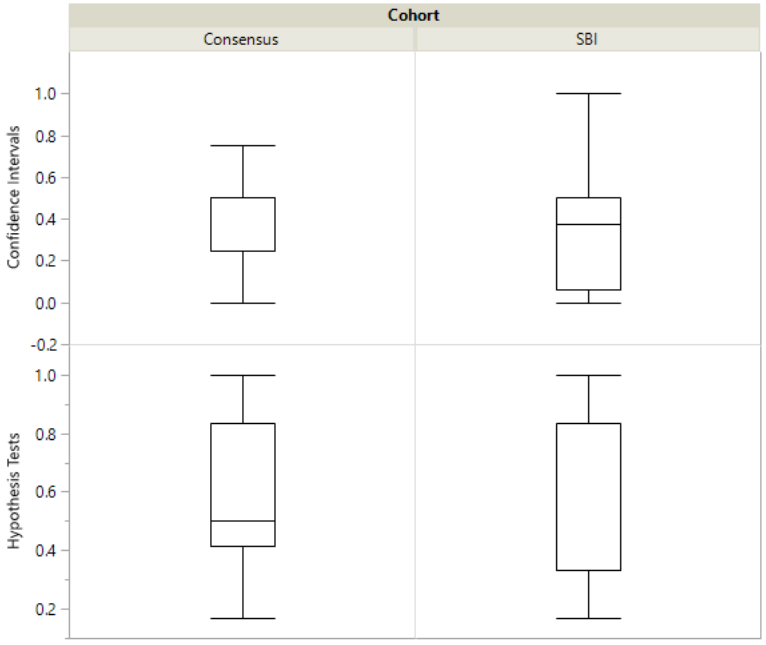

Figure 3: Final exam average proportion correct

Table 12: Comparison of final exam average percent correct by topic

		Confidence Intervals		Hypothesis Tests	
	n	Mean (SD)	95% CI	Mean (SD)	95% CI
Consensus	33	40.9 (22.4)	(33.0, 48.8)	61.1 (26.9)	(51.6, 70.6)
SBI	12	37.5 (31.1)	(17.8, 57.2)	48.6 (29.7)	(29.7, 67.4)

4.2.2 Final Exam Question by Question Comparison for Confidence Intervals

Question 23 asked students to identify the relationship between confidence level and width of the confidence interval, and a similar percentage of students in each curriculum answered this question correctly (Table 13). In addition, a similar percentage of students in each curriculum selected choice D, which was the lowest confidence level given and might be considered the second best response. That is, if students understand that width is affected by confidence level they should be able to reason that the correct answer is either the highest

confidence level or the lowest. Nonetheless, five students (15.2%) in the consensus cohort and two students (16.7%) in the SBI group selected one of the "middle" confidence levels as their answer.

Question 25 asked students to identify the relationship between sample size and width, and a similar percentage of students in each curriculum answered this question correctly (Table 13), although it was only about one quarter of each group. Choice B, which demonstrates the misconception that a larger sample size leads to a wider confidence interval, was the most popular choice in both curriculum groups, selected by 45.5% of consensus students and 75.0% of SBI students. It is interesting to note that about one-fourth of the students in the consensus curriculum selected choice D, in which the midpoint of the confidence interval is different (incorrect) but Elsa's interval is narrower (correct). Also, none of the students in the SBI curriculum chose either option A or D.

About twice as many students in the consensus curriculum as compared to the SBI curriculum selected the correct answer for question 28, addressing the purpose of confidence intervals (Table 13). It is troubling to note that almost 42% of students in the SBI curriculum, as well as 24% of students in the consensus curriculum, indicated that confidence intervals are used to estimate the value of the p-value. In addition, 19.1% of students in the consensus cohort selected the standard error as the target of estimation for a confidence interval. Both groups saw two similar questions in advance of the final exam, although these questions asked students to recognize that the sample statistic is always contained in the confidence interval. Perhaps this should have prompted students to eliminate the sample proportion as the target of estimation for the confidence interval, which was chosen by 18.2% of consensus students and 33.3% of SBI

students. It is worth noting that the prior questions came much closer in time to the final exam for the consensus students (about 1.5 weeks) than the SBI students (about 5.5 weeks).

Understanding the meaning of the confidence level (question 29) is a difficult topic for students. One common misinterpretation is to describe the confidence level as the probability that the parameter is contained in the confidence interval. Students were repeatedly reminded that this interpretation is incorrect, and the interpretation was not given as an answer option in this question. A larger percentage of SBI students, 41.7% compared to 24.2% in the consensus group, correctly interpreted the confidence level as the percentage of confidence intervals that will succeed in capturing the parameter (Table 13). The SBI students had the exact same question on their second unit exam, although this took place about 5.5 weeks before the final exam. It is interesting to note that 33.3% of students in the consensus cohort and 16.7% of students in the SBI curriculum interpreted the confidence level to correspond to the percentage of intervals that would contain the sample statistic. This seems to contradict the results of the previous question, where the percentages were approximately opposite for the two curriculum groups in concluding that the purpose of the confidence interval is to estimate the value of the sample proportion. The question statement indicates that each student is constructing "a 95% confidence interval for the mean height of all DCC students", once again suggesting that the purpose of the confidence interval is to estimate the parameter as opposed to the statistic. In addition, questions 28 and 29 were next to each other on the final exam, so students could have reevaluated their answer to question 28 after seeing question 29. A similar proportion of students in each cohort (about 42%) interpreted the confidence level to be the percentage of population values that would be contained in the confidence intervals. This may indicate confusion with the empirical rule, in which 95% of data values are within 2 standard deviations

of the mean when the population is normally distributed. Since the variable in question was height, it is perhaps not entirely unreasonable to imagine that the variable is normally distributed in the population. Finally, none of the students picked choice D, in which the confidence level is interpreted as the percentage of constructed confidence intervals that would be identical.

Table 13: Final exam confidence interval subtopics

	Consensus (n = 33)	SBI (n = 12)
	Percent Correct	
Relationship between confidence level and width	63.6	58.3
Relationship between sample size and width	27.3	25.0
Purpose of confidence intervals	48.5	25.0
Interpretation of confidence level	24.2	41.7

4.2.3 Final Exam Question by Question Comparison for Hypothesis Tests

Results for the questions addressing the relationship between p-value and *z*-score and whether a p-value can be negative (questions 7 and 16, respectively) were similar for both curriculum groups (Table 14). These were both True/False questions.

Although students in the SBI curriculum worked with hypothesis tests throughout the semester starting in the second week, only one-third correctly stated on question 20 that the hypotheses are always about the parameter (Table 14). By contrast, about half (51.5%) of the consensus group correctly identified this statement (Table 14). Only one student (3.0%) in the consensus cohort, compared to four (one-third of) SBI students, said that the hypotheses in a test of significance are always about the statistic only. Another one-third of SBI students and 27.3% of consensus students stated that the hypotheses are always about both the statistic and the parameter. Finally, 18.2% of students in the consensus group stated that the hypotheses are

sometimes about the statistic and sometimes about the parameter, whereas no SBI students chose this option.

Question 12 asked students to identify the definition of p-value, and this was answered correctly by about three-quarters of students in the consensus cohort, compared to 58.3% of SBI students (Table 14). One common misconception of p-value is that it is the probability that the null hypothesis is true. This option was selected by one student (3.0%) in the consensus group and two students (16.7%) in the SBI group. In addition, these same proportions of students in each cohort selected "The probability that the alternative hypothesis is true" as the definition of the p-value, which is another common misconception. Only one student (8.3%) in the SBI cohort and six students (18.2%) in the consensus group selected choice A, which gave the definition of p-value as the probability of getting a result at least as extreme as the sample statistic assuming the alternative hypothesis is true. The consensus students had two similar questions, one on a practice exam and one on a unit exam, about 1.5 weeks before the final exam. The SBI group had a similar question on their first unit exam, which was 11.5 weeks before the final exam. However, hypothesis tests were run and p-values were determined throughout the semester in the SBI class.

In addition to understanding the definition of the p-value, it is important for students to be able to apply this definition in the context of a study. Question 30 addressed this learning goal by providing a context that was local, wherein a county legislator wanted to find out whether a majority of voters supported a proposed tax to continue work on the Dutchess County Rail Trail. Choice B, which stated that the p-value (0.08) was extreme and that the test proved that the null hypothesis was true, was chosen by only one student, who was in the consensus cohort. The other three answer choices required students to decide whether the p-value should be doubled for

a two-sided test (choice A) or whether the one-sided p-value was surprising (choice D) or not surprising (choice C). Students were also given a *z*-score of 1.40 to accompany the p-value. While over half (54.5%) of consensus students and one-third of SBI students selected the correct answer (Table 14), another quarter (24.2%) of consensus students and one-third of SBI students chose the option for a two-sided test. While this answer choice correctly stated that the result was not surprising, it appears that students did not recognize that the word "majority" in the question statement referred to a one-sided alternative. Another 18.2% of the consensus group and one-third of the SBI group stated that a p-value of 0.08 indicated a surprising result that could not easily happen by chance. Since the three most popular answer choices in both groups began with the phrase "If the null hypothesis is true", it would be interesting to know whether students went back to the question about the definition of p-value to reassess their answers after reading this question. It is also worth noting that the consensus students saw five similar questions prior to the final exam, four of which were on the practice for the fourth unit exam, and one of which was on the fourth unit exam. Students in the SBI cohort saw two similar questions, one on the practice for the first unit exam and one on the first unit exam. Both of these questions involved the same context, and the wording of the answer choices was not as similar to the final exam question as it was for the consensus students.

Question 31 involved the general interpretation of the p-value in terms of the null hypothesis and stated that the p-value under consideration was "very large". The two least popular answer choices in both cohorts involved the word "proven", which students were warned against multiple times during the semester (recall that in the previous paragraph, the answer choice involving the word "proves" was chosen by only one student). About 54.5% of consensus students and 41.7% of SBI students correctly stated that a very large p-value indicates

that the null hypothesis is plausible (Table 14). On the other hand, 36.4% of the consensus group and 41.7% of the SBI cohort stated that a very large p-value indicates that there is strong evidence for the alternative hypothesis.

Table 14: Final exam hypothesis test subtopics

	Consensus (n = 33)	SBI (n = 12)
	Percent Correct	
Relationship between p-value and *z*-score	60.6	58.3
Definition of p-value	75.8	58.3
Whether a p-value can be negative	69.7	66.7
Hypotheses are statements about parameters	51.5	33.3
Interpretation of p-value (in context)	54.5	33.3
Interpretation of p-value (no context)	54.5	41.7

CHAPTER 5

DISCUSSION

5.1 Summary of Findings

Overall, the data do not provide evidence that students following a curriculum using simulation-based inference (SBI) learn more about statistical inference than students following a more traditional curriculum. The small sample sizes, however, especially for the SBI cohort, make it difficult to detect a statistically significant difference in learning. In addition, although there was at least moderate evidence of a significant increase in general learning from pretest to posttest for each curriculum, it is disappointing to note that only three students scored higher than 70% on the R-CAOS posttest: two from the consensus curriculum and one from the SBI cohort. These students had the largest increase from pretest to posttest, with a gain of nearly forty percentage points for each student.

Overall, students did reasonably well at identifying the correct interpretation of both a confidence interval and a p-value. On the R-CAOS posttest, 85.7% of SBI students and 61.5% of consensus students identified a correct interpretation of a confidence interval. Even so, a large proportion of students in both curriculum groups also stated that incorrect interpretations of confidence intervals and p-values were valid. These percentages are in line with the studies by delMas et al. (2007) and Tintle et al. (2011). One bright spot in this research project is with the misconception that the p-value is the probability that the alternative hypothesis is true. A relatively large proportion of SBI students (71.4%) were able to identify this interpretation as invalid on the R-CAOS posttest, compared to 33.3% of the consensus cohort. In the national

data analyzed by delMas et al. (2007), only 52.7% of students stated that this interpretation was invalid on the CAOS posttest. In the sample collected by Tintle et al. (2011), only 44.6% of SBI students and 47.7% of consensus students correctly identified this interpretation as invalid on the CAOS posttest.

As is the case with the R-CAOS data, the average percent correct on confidence interval and hypothesis test questions on the final exam (Table 12) is also disappointing, particularly in light of the fact that students in both groups had encountered similar questions on quizzes, practice exams, and unit exams throughout the semester. It is worth noting that, for almost every subtopic, students in the consensus curriculum were given a greater number of similar questions in advance of the final exam than SBI students (Table 2). In addition, questions for the consensus cohort came much closer in time to the final exam than for the SBI group. Although students following the SBI curriculum worked with p-values all semester, the focus was more on evaluating strength of evidence against the null hypothesis provided by the p-value rather than interpreting the p-value in terms of its definition. The question on the final exam that involved the interpretation of p-value in context was written in terms of the definition of p-value rather than strength of evidence. Given this background, it is perhaps not surprising that a greater percentage of consensus students answered this question correctly on the final exam (Table 14).

5.2 Limitations of the Study

The biggest limitation in this research project is the small sizes of the samples, especially the SBI cohort. As pointed out in Tintle et al. (2011), differences in administration, incentives offered, and demographics between their samples and the national sample analyzed by delMas et al. (2007) may have impacted results and limited the scope of conclusions that could be drawn from their analysis. These considerations, as well as the small sample sizes, are also factors in

this research study. In addition, there were five students who were missing data on the R-CAOS assessment, which resulted in averages that were taken over fewer questions. If the sample sizes were larger, all students with incomplete data could have been eliminated from the analyses.

It is also important to point out that simulation methods were used briefly in the consensus classes when sampling distributions were covered. Specifically, students were given guided instructions that enabled them to create a hypothetical population in StatCrunch from which to draw simulated samples and consider the sampling distribution of statistics for a single categorical variable. The main purpose of simulation in this case was to help students understand the difference between the population distribution, sample distribution, and sampling distribution. Students in the consensus cohort did not use simulation to find p-values or to create confidence intervals, as SBI students did. And, as mentioned previously, consensus students used simulation briefly and only in the last few weeks of the semester, while the SBI students used simulation all semester long. Still, it is possible that the use of simulation did enhance the understanding of consensus students even though it was a much smaller part of the curriculum. In addition, it was the researcher's first time teaching the SBI curriculum. Since it was the second time teaching the consensus curriculum with the Gould and Ryan (2016) text, the researcher had more materials prepared based on the Fall 2015 semester. In addition, the researcher already had extensive experience incorporating activities into a traditional introductory statistics course and knew which concepts and topics required focused attention. The SBI curriculum as presented in the Tintle et al. (2016) text is a radical departure from the traditional Stat 101 course and did not provide the opportunity for incorporating additional activities and teaching strategies in the first implementation. For these reasons, it is perhaps

unfair to compare the two curricula until more experience is gained teaching with the SBI curriculum.

As mentioned in Section 5.1, the uncontrolled testing environment for the R-CAOS assessment may also have affected results. On one hand, students may have felt that they did not need to take the test too seriously since they would get a 100% homework grade for completing the assessment regardless of their score. On the other hand, there also was no incentive for students to try to get help from their textbooks or the Internet or other sources. Students were told during the IRB process that their data would be used for research purposes and that they should take the test in a quiet and distraction-free environment and to answer the questions to the best of their ability. In the future, it may be worthwhile to administer both the pretest and posttest in class in a more controlled environment, and to tie the score on the R-CAOS posttest to the student's grade in a more meaningful way. For example, the R-CAOS posttest score could count as a percentage of the final exam grade, or extra credit could be given on a sliding scale based on proportion of questions answered correctly.

5.3 Future Directions

This research study has given rise to ideas for future research as well as implications for teaching, which are described in the following two sections.

5.3.1 Future Directions for Research

There are many avenues for future research regarding the use of simulation-based inference to teach introductory statistics. It has been conjectured that teaching hypothesis testing throughout the semester starting in the first or second week will give students a better understanding of the core ideas and logic of inference than they would get in a more traditional curriculum, where inference is typically covered in the last third of the semester (see, for

example, Cobb, 2007; Garfield et al., 2012; Holcomb et al., 2010; Malone et al., 2010; Roy et al., 2014; Tintle et al., 2011). It would be interesting to study the question of how understanding of statistical inference develops over time. For example, students could be given similar questions addressing specific concepts throughout the semester to see whether the proportion of students that demonstrate understanding increases during the course of the semester.

The importance of considering attitudes in statistics education has been a topic of concern among some researchers (e.g., Gal, Ginsburg, & Schau, 1997; Gal & Ginsburg, 1994; Ramirez, Schau, & Emmioglu, 2012; Schau, Millar, & Petocz, 2012). Some research has been conducted to compare attitudes among students following a traditional curriculum to those following a randomization-based curriculum (Swanson, VanderStoep, & Tintle, 2014). It was originally intended that in this research study, student attitudes would be compared between the consensus and SBI curriculum groups. This part of the project was abandoned, however, due to the small sample sizes. It is hoped that future research will include assessment of student attitudes from a randomization-based curriculum. Since most, if not all, sections of MAT118 at DCC will be taught using SBI in the future, it is unclear whether comparisons will be made between students following a consensus curriculum and students using the SBI approach, but comparisons could be made to results found in other studies (e.g., Schau & Emmioglu, 2012).

Another issue that has been noted by instructors at Dutchess Community College is the level of reading ability that is required to be successful with the SBI curriculum. The problems students encounter in the SBI course, since they are mostly based on real research studies, necessarily require higher level reading comprehension skills. Students can watch videos that cover the main points in the textbook, thus reducing the need to read the text. What is uncertain, however, is whether students recognize that they need to watch the videos with a goal of

understanding the ideas being explained, as opposed to the way they watch television or other entertainment videos. It would be useful to conduct research in this area to determine whether students with higher reading levels, perhaps as measured by standardized tests such as the ACT or SAT, tend to be more successful in the SBI course. Another question that could be addressed by research is whether students who have already completed a 100-level English course tend to be more successful in the SBI course compared to students who are in a remedial (0-level) English course while taking introductory statistics, or students who are taking their first college level English course while taking introductory statistics.

5.3.2 Implications for Teaching

There are also implications for teaching that follow from this research project. Regarding reading ability discussed in the previous section, one DCC instructor has suggested creating "notes guides" for students to complete while watching the videos and/or reading the textbook. This might help students focus on the important concepts in a section and be more prepared to participate in associated activities. The department may also decide to require a reading and/or English pre- or co-requisite before or while taking MAT118.

As discussed in Section 5.1, although p-values are used throughout the entire semester in the SBI curriculum, much more emphasis is placed on evaluating strength of evidence against the null hypothesis than on the interpretation of the definition of the p-value. In future versions of MAT118 using SBI, it is recommended that students spend more time interpreting p-values in the context of each study that they encounter.

Faculty at DCC are also currently involved in a remedial redesign project, one aspect of which involves the introductory statistics course. The prerequisite for DCC's MAT118 course is MAT092, a remedial level quantitative literacy course. There is agreement among the DCC

mathematics faculty that an algebra prerequisite is not appropriate for MAT118, especially if MAT118 is taught using SBI. Also, there are several topics in MAT092 that are not necessarily helpful in preparing students for MAT118. The researcher's experience with SBI indicates that students should be very comfortable with decimals, fractions, and percentages in order to succeed in the SBI version of MAT118. In addition, it is often the case that DCC students (and, indeed, community college students in general) are not prepared for the amount of work and level of responsibility and independence that are required in a college course. The DCC mathematics faculty are now considering a co-requisite model for MAT092 wherein students will be able to complete both MAT092 and MAT118 in a single semester. Although the details have yet to be worked out, the main idea is to give students the necessary content, reading, and study skills support in MAT092 while they are taking MAT118. It is hoped that this model will help more students succeed in MAT118, while allowing them to complete their college course and prerequisite course in a single semester.

Finally, the researcher would like to incorporate a semester project into the MAT118 course. The researcher has attended many conference workshops that suggest students get much more out of the introductory statistics course when they complete a project of this kind. The revised GAISE guidelines (GAISE College Report ASA Revision Committee, 2016) also recommend that a project is one way to incorporate the investigative process into Stat 101. Requiring students to develop a research question, collect data, and analyze that data to answer the research question may help students to master many of the main ideas of MAT1118 and to confront misconceptions that remain after typical classroom instruction. This may also help students to see the usefulness and relevance of statistics to their lives.

While the results from this study are not robust due to sample size, the DCC mathematics faculty will benefit from the research as they continue to refine the introductory statistics course. Data will be collected moving forward so that student learning can be better understood, not just with respect to inference but all the important topics in a first course in statistics. A cycle of continuous improvement has now begun that can be implemented for the foreseeable future. In addition, research conducted at DCC may be used to contribute to the body of statistics education research.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education: College Report*. Alexandria, VA: American Statistical Association. [Online: http://www.amstat.org/education/gaise/]

ARTIST (2006). Assessment Resource Tools for Improving Statistical Thinking (ARTIST). [Online: https://apps3.cehd.umn.edu/artist/]

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1-2).

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.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic visualizations and the randomization test. *Technology Innovations in Statistics Education*, *7*(2).

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3).

Carver, R. (2011). Introductory statistics unconstrained by computability: A new cobb salad. *Technology Innovations in Statistics Education*, *5*(1).

Castro Sotos, A., Vanhoof, S., Noorgate, 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.

CATALST (2012). NSF Award DUE-0814433. Change agents for teaching and learning statistics (CATALST). [Online: http://www.tc.umn.edu/~catalst/]

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.

Cobb, G. (1992). Teaching statistics. In Lynn A. Steen (Ed.), *Heeding the Call for Change: Suggestions for Curricular Action* (MAA Notes No. 22, pp. 3-43). Washington, DC: Mathematical Association of America.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). [Online: http://escholarship.org/uc/item/6hb3k0nz]

delMas, R. C., Garfield J., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3).

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

Fidler, F. (2006). Should psychology abandon *p*-values and teach CIs instead? Evidence-based reforms in statistics education. In A. J. Rossman & B. L. Chance (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics* (ICOTS-7). Voorburg, The Netherlands: International Statistical Institute. [Online: https://www.stat.auckland.ac.nz/~iase/publications/17/5E4_FIDL.pdf]

GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in Statistics Education College Report 2016". [Online: http://www.amstat.org/education/gaise]

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2). [Online: https://ww2.amstat.org/publications/jse/v2n2/gal.html]

Gal, I., Ginsburg, L., & Schau, C. (1997). Monitoring attitudes and beliefs in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 37–51). Amsterdam, The Netherlands: IOS Press. [Online: https://www.stat.auckland.ac.nz/~iase/publications/assessbk/chapter04.pdf]

Garfield, J. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*(7).

Gould, R. & Ryan, C. (2016). *Introductory statistics: Exploring the world through data*, Second Edition. Boston: Pearson Higher Ed.

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

Holcomb, J., Chance, B., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference. In C. Reading (Ed.), *Data and Context in Statistics Education: Towards an Evidence-based Society. Proceedings of the Eighth International Conference on the Teaching of Statistics* (ICOTS-8). Voorburg, The Netherlands: International Statistical Institute. [Online: https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_5F1_CHANCE.pdf]

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Data and Context in Statistics Education: Towards an Evidence-based Society. Proceedings of the Eighth International Conference on the Teaching of Statistics* (ICOTS-8). Voorburg, The Netherlands: International Statistical Institute. [Online: https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8D1_HOLCOMB.pdf]

Jones, P., Lipson, K., & Phillips, B. (1994). A role for computer intensive methods in introducing statistical significance. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (IASE). Perugia, Italy: University of Perugia. [Online: http://www.stat.auckland.ac.nz/~iase/publications/proc1993/255-263rec.pdf]

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Developing a Statistically Literate Society: Proceeding of the Sixth International Conference on the Teaching of Statistics* (ICOTS-6). Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/1/6c1_lips.pdf]

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a Web-based computer simulation. *Proceeding of the IASE Satellite Conference on Statistics Education and the Internet*. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/6/lipson.pdf]

Konold, C. (1994). Understanding probability and statistical inference through resampling. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (IASE). Perugia, Italy: University of Perugia. [Online: http://www.stat.auckland.ac.nz/~iase/publications/proc1993/199-211rec.pdf]

Konold, C. & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4).

Lock, R., Lock, P., Morgan, K., Lock, E., & Lock, D. (2013). *Statistics: Unlocking the Power of Data*. Hoboken, NJ: John Wiley & Sons Incorporated.

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1).

Maurer, K. & Lock, D. (2016). Comparison of learning outcomes for simulation-based and traditional inference curricula in a designed educational experiment. *Technology Innovations in Statistics Education, 9*(1).

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1).

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistics Review, 65*, 123-165.

Ramirez, C., Schau, C., & Emmioglu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal*, *11*(2).

Rossman, A. J. & Chance, B. L. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *College Mathematics Journal*, *30*(4).

Rossman, A. J., Chance, B. L., Cobb, G., & Holcomb, J. (2008). NSF/CCLI/DUE-0633349. Concepts of statistical inference: A randomization-based curriculum. [Online: http://statweb.calpoly.edu/csi]

Roy, S., Rossman, A., Chance, B., Cobb, G., VanderStoep, J., Tintle, N., & Swanson, T. (2014). Using simulation/randomization to introduce p-value in week 1. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education: Proceedings of the Ninth International Conference on the Teaching of Statistics* (ICOTS-9). Voorburg, The Netherlands: International Statistical Institute. [Online: http://icots.info/9/proceedings/pdfs/ICOTS9_4A2_ROY.pdf]

Saldanha, L. A. & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3).

Schau, C. (2003). Survey of Attitudes Toward Statistics (SATS-36). [Online: http://evaluationandstatistics.com/]

Schau, C. & Emmioglu, E. (2012). Do introductory statistics courses in the United States improve students' attitudes? *Statistics Education Research Journal*, *11*(2).

Schau, C., Millar, M., & Petocz, P. (2012). Research on attitudes towards statistics. *Statistics Education Research Journal*, *11*(2).

Simon, J. L. (1994). What some puzzling problems teach about the theory of simulation and the use of resampling. *The American Statistician*, *48*(4).

Simon, J. L., Atkinson, D., & Shevokas, C. (1976). Probability and statistics: Experimental results of a radically different teaching method. *American Mathematical Monthly*, *83*(9).

Swanson, T., VanderStoep, J., & Tintle, N. (2014). Student attitudes toward statistics from a randomization-based curriculum. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education: Proceedings of the Ninth International Conference on the Teaching of Statistics* (ICOTS-9). Voorburg, The Netherlands: International Statistical Institute. [Online: http://icots.info/icots/9/proceedings/pdfs/ICOTS9_1F1_SWANSON.pdf]

Tintle, N., Chance, B. L., Cobb, G., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: John Wiley & Sons Incorporated.

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). Quantitative evidence for the use simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education: Proceedings of the Ninth International Conference on the Teaching of Statistics* (ICOTS-9). Voorburg, The Netherlands: International Statistical Institute. [Online: http://icots.info/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf]

Tintle, N., Swanson, T., & VanderStoep, J. (2009). *An Active Approach to Statistical Inference*, Preliminary Edition, Holland, MI: Hope College Publishing.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1). [Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ11%281%29_Tintle.pdf]

Tintle, N., VanderStoep, J., Holmes, V-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1). [Online: http://www.amstat.org/publications/jse/v19n1/tintle.pdf]

Well, A.D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*, 289–312.

APPENDIX A

ASSESSMENT QUESTIONS

Responses in **bold** are correct responses.

Table 15: R-CAOS questions

Subtopic	Question
Relationship between sample size and variability	34. Suppose at a large university, 15% of the students are left-handed. Sam plans to take a random sample of 100 students and ask whether or not the student is left-handed. Kerry plans to ask a random sample of 50 students whether or not the student is left-handed. Who, Sam or Kerry, is more likely to find more than 25% of their sample is left-handed? A. Sam because a larger sample is more likely to have more left-handed students. B. Kerry because a smaller sample is more likely to have more left-handed students. **C. Kerry because there is more variability in the sample proportions among smaller samples.** D. Sam because there is more variability in the sample proportions among larger samples. E. Both have the same chance because both are planning to select a random sample from a population in which 15% are left-handed.

Relationship between sample size and variability

44. When manufactured, pennies need a beveled edge (slightly angled) to help pop them out of the press. For this reason, it has been conjectured that spinning a penny on its edge is more likely to land with the tail side up than with the head side up. Suppose you investigate by spinning a penny 15 times (put it on its edge and flick it to spin on its own) and that you find that the penny lands with the tail side up 13 times. You determine that if a spun penny is equally likely to land tails or heads, then the probability of 13 or more tails in 15 coin spins is 0.004. For which outcome would you be more convinced that this coin is more likely to land tails when spun?

A. 13 tails in 15 coin spins
B. 130 tails in 150 coin spins
C. They are equally convincing because 13/15 = 130/150

Interpretation of confidence interval

A high school statistics class wants to estimate the average cookie weight of a generic brand of chocolate chip cookies. They collect a random sample of 50 cookies from the manufacturing process and obtain the weight (in grams) for each cookie. Based on their data, the 95% confidence interval for the average weight per cookie is 25.65 to 26.35 grams. For each of the following three statements, indicate whether it is a valid or invalid conclusion.

Interpretation of confidence interval

18. We can infer with 95% confidence that a randomly selected cookie manufactured for this generic brand will weigh between 25.65 to 26.35 grams.

A. Valid **B. Invalid**

Interpretation of confidence interval

19. We can infer with 95% confidence that mean weight of all cookies manufactured for this generic brand is between 25.65 and 26.35 grams.

A. Valid B. Invalid

Interpretation of confidence interval

20. We can infer with 95% confidence that the average weight for 50 cookies randomly selected from those manufactured for this generic brand will be between 25.65 and 26.35 grams.

A. Valid **B. Invalid**

46. Suppose your teacher believes the confidence interval found in the previous question is too wide. She wants to know what could have been done to produce a narrower confidence interval and therefore a more precise estimate of the mean foot length for students at this university. For each suggestion below, answer True, False, or Can't Tell for whether this change would produce a narrower confidence interval.

Relationship between sample size and width

Increase the sample size to 150

A. Yes B. No C. Can't Tell

Relationship between confidence level and width

Increase the confidence level to 99%

A. Yes **B. No** C. Can't Tell

Small p-value is desirable

27. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p-value would she want to obtain?

A. The magnitude of a p-value has no impact on statistical significance.
B. A large p-value
C. A small p-value

Interpretation of p-value (in context)

A research article reports the results of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a p-value of 0.04 in the analysis section. Indicate whether the following interpretations are valid or invalid interpretations of this p-value.

Interpretation of p-value (in context)

28. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.

A. Valid **B. Invalid**

Interpretation of p-value (in context)	29.	We conclude that the new drug is effective because the results they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective. **A. Valid** B. Invalid
Interpretation of p-value (in context)	30.	We conclude that the new drug is effective because there is only a 4% chance that it's not. A. Valid **B. Invalid**
Interpretation of p-value (in context)	43.	When manufactured, pennies need a beveled edge (slightly angled) to help pop them out of the press. For this reason, it has been conjectured that spinning a penny on its edge is more likely to land with the tail side up than with the head side up. Suppose you investigate by spinning a penny 15 times (put it on its edge and flick it to spin on its own) and that you find that the penny lands with the tail side up 13 times. You determine that if a spun penny is equally likely to land tails or heads, then the probability of 13 or more tails in 15 coin spins is 0.004. What does this analysis tell you about whether this penny is more likely to land tails than heads if spun a large number of times? A. Getting 13 tails in 15 spins likely happened just by chance and therefore this penny has a 50-50 chance to land tails when spun a large number of times. **B. There is strong evidence that this coin is more likely to land tails than heads if spun a large number of times.** C. These results prove that this penny is more likely to land tails than heads when spun a large number of times. D. Nothing, spinning the penny only 15 times does not produce conclusive results either way.

Table 16: Final exam questions

Subtopic	Question
Relationship between confidence level and width	23. A confidence interval is constructed to estimate the population proportion of consumers who will visit a store because of a sale. Which confidence interval would be the widest? A. 90% **B. 99%** C. 95% D. 85%
Relationship between sample size and width	25. Suppose that Elsa and Frank determine confidence intervals using the same confidence level, based on the same sample proportion. Elsa uses a larger sample size than Frank. How will the midpoint and width of the confidence intervals compare? A. Same midpoint and same width. B. Same midpoint, Elsa has a wider interval. **C. Same midpoint, Frank has a wider interval.** D. Elsa has a larger midpoint, narrower interval.
Purpose of confidence intervals	28. We use confidence intervals to estimate the value of _____ . **A. the population proportion** C. the p-value B. the sample proportion D. the standard error
Interpretation of confidence level	29. Suppose we have a collection of the heights of all students at DCC. Each of the 250 people taking MAT118 randomly takes a sample of 40 of these heights and constructs a 95% confidence interval for the mean height of all DCC students. Which of the following statements about the confidence intervals is most accurate? A. About 95% of the heights of all DCC students will be contained in these intervals. B. About 95% of the time, a student's sample mean height will be contained in his or her interval. **C. About 95% of the intervals will contain the population mean height.** D. About 95% of the intervals will be identical.
Relationship between p-value and *z*-score	7. A small (close to 0) p-value goes with a large (far from 0) standardized (*z*) statistic. **A. True** B. False

Definition of p-value

12. Which of the following is the correct definition of the p-value?

A. The probability of getting a result at least as extreme as the sample statistic if the alternative hypothesis is true.
B. The probability that the null hypothesis is true.
C. The probability of getting a result at least as extreme as the sample statistic if the null hypothesis is true.
D. The probability that the alternative hypothesis is true.

Whether a p-value can be negative

16. A p-value can be negative.

A. True **B. False**

Hypotheses are statements about parameters

20. When stating null and alternative hypotheses, the hypotheses are _____ .

A. always about the parameter only
B. always about the statistic only
C. always about both the statistic and the parameter
D. sometimes about the statistic and sometimes about the parameter

Interpretation of p-value (in context)

30. A Dutchess County legislator conducts a test of significance to determine whether a majority of voters support a proposed tax to add to the Dutchess County Rail Trail. The calculated standardized statistic is $z = 1.40$ with an associated p-value of 0.081. Choose the conclusion that provides the best interpretation for the p-value at a significance level of 0.05.

A. If the null hypothesis is true, then the probability of getting a standardized statistic that is at least as extreme as 1.40 is 0.081. The result should be doubled for a two-sided test. This result is not surprising and could easily happen by chance.
B. The p-value should be considered extreme; therefore the hypothesis test proves that the null hypothesis is true.
C. If the null hypothesis is true, then the probability of getting a standardized statistic at least as large as 1.40 is 0.081. This result is not surprising and could easily happen by chance.
D. If the null hypothesis is true, then the probability of getting a standardized statistic at least as extreme as 1.40 is 0.081. This result is surprising and could not easily happen by chance.

Interpretation of p-value (no context)

31. When we get a p-value that is very large we may conclude that _____ .

A. the null hypothesis has been proven to be true
B. there is strong evidence for the alternative hypothesis
C. the null hypothesis is plausible
D. the alternative hypothesis has been proven to be false

Table 17: Similar questions given prior to final exam

Subtopic	Cohort Instrument	Question
Relationship between confidence level and width	Consensus Exam 4	From a random sample of workers at a large corporation you find that 58% of 200 went on a vacation last year away from home for at least a week. An approximate 95% confidence interval is (0.50, 0.66). Which of the following statements is correct concerning this situation? A. If the sample size is 500 instead of 200, the confidence interval will be wider. B. A maximum of 66% of all coworkers went on a vacation last year away from home for at least a week. **C. If the confidence level were changed from 95% to 99%, the confidence interval would become wider.** D. If the confidence level were changed from 95% to 90%, the confidence interval would become wider.
Relationship between confidence level and width	Consensus Exam 4 SBI Exam 2	A student was interested in estimating what proportion of San Luis Obispo (SLO) residents dine at restaurants at least once a week. She selected a random sample of adult residents of SLO and asked each participant whether he or she dines at restaurants at least once a week. She used her data to find a 95% confidence interval for the proportion of all SLO adults who dine out at least once a week. The confidence interval was (0.38, 0.44). The 99% confidence interval based on the same data would be _____ . A. narrower, because the more confident we are the narrower the interval **B. wider, because to be more confident we need to widen the interval** C. More information is needed to answer this question.
Relationship between confidence level and width	SBI Exam 2	A confidence interval is constructed to estimate a population parameter. Which confidence level would produce the widest interval? **A. 99%** B. 90% C. 95% D. 85%

Relationship between confidence level and width Relationship between sample size and width	Consensus Practice Exam 4	A pollster took a random sample of 100 students from a large university and computed a 95% confidence interval to estimate the percentage of students who were planning to vote in the upcoming election. The pollster felt that the confidence interval was too wide to provide a precise estimate of the population parameter. What could the pollster have done to produce a narrower confidence interval that would give a more precise estimate of the percentage of all university students who plan to vote in the upcoming election? A. Increase the sample size to 150 B. Increase the confidence level to 99% **C. Both A. and B.** D. None of the above
Relationship between sample size and width	Consensus Practice Exam 4	Two different samples will be taken from the same population where the population proportion is unknown. The first sample will have 25 data values, and the second sample will have 64 data values. A 95% confidence interval will be constructed for each sample to estimate the population proportion. Which confidence interval would you expect to have greater precision (a smaller width) for estimating the population mean? **A. I expect the confidence interval based on the sample of 64 data values to be more precise.** B. I expect both confidence intervals to have the same precision. C. I expect the confidence interval based on the sample of 25 data values to be more precise.
Relationship between sample size and width	Consensus Practice Exam 4	When constructing a confidence interval, a larger sample size will _____ the margin of error and the confidence interval will be _____ . A. increase; narrower B. increase; wider C. **decrease; narrower** D. decrease; wider

Relationship between sample size and width	SBI Exam 2	For the same sample proportion and confidence level, what happens to the width of a confidence interval if the sample size is increased? A. The width increases **B. The width decreases** C. The width stays the same D. It is impossible to say
Purpose of confidence intervals	Consensus Practice Exam 4	A 95% confidence interval is computed to estimate the proportion of adults who earn more than $50000 per year. Which of the following will definitely be within the limits of this confidence interval? A. The population proportion **B. The sample proportion** C. The standard deviation of the sample proportion D. None of the above
Purpose of confidence intervals	Consensus Exam 4 SBI Quiz #5 Exam 2	Suppose a 95% confidence interval for a population proportion is found using StatCrunch ["StatCrunch" was changed to "the 2 SD method" on SBI evaluations]. Which of the following will *definitely* be contained in that interval? A. The population proportion **B. The sample proportion** C. The p-value D. All of the above
Interpretation of confidence level	Consensus Practice Exam 4	Each of the 110 students in a statistics class selects a different random sample of 35 quiz scores from a population of 5000 scores they are given. Scores can range from 0 to 100. Using their data, each student constructs a 90% confidence interval for the population proportion of scores that are above 70. Which of the following conclusions is correct? A. About 10% of the sample proportions will not be included in the confidence intervals. **B. About 90% of the confidence intervals will contain the population proportion.** C. It is probable that 90% of the confidence intervals will be identical. D. About 10% of the raw scores in the samples will not be found in these confidence intervals.

Interpretation of confidence level	Consensus Exam 4	From a random sample of workers at a large corporation you find that 58% of 200 went on a vacation last year away from home for at least a week. An approximate 95% confidence interval is (0.50, 0.66). Which of the following statements is a correct interpretation of the confidence interval? A. 95% of the coworkers fall in the interval (0.50, 0.66). **B. We are 95% confident that the proportion of all coworkers who went on a vacation last year away from home for at least a week is between 50% and 66%.** C. There is a 95% chance that a randomly selected coworker has gone on a vacation last year away from home for at least a week. D. We are 95% confident that between 50% and 66% of the samples will have a proportion near 58%.
Interpretation of confidence level	SBI Exam 2	Suppose we have a collection of the heights of all students at DCC. Each of the 250 people taking statistics randomly selects a sample of 40 DCC student heights and constructs a 95% confidence interval for mean height of all students at DCC. Which of the following statements about the confidence intervals is most accurate? A. About 95% of the heights of all students at DCC will be contained in these intervals. B. About 95% of the time, a student's sample mean height will be contained in his or her interval. **C. About 95% of the intervals will contain the population mean height.** D. There is a 95% chance that any one of the confidence intervals contains the population mean height.
Relationship between p-value and *z*-score	SBI Practice Exam 1	A small (close to 0) p-value goes with a large (far from 0) *z* statistic. **A. True** B. False

Relationship between p-value and *z*-score	SBI Practice Exam 1	Suppose you are testing H_0: $\pi = 0.5$ versus H_a: $\pi < 0.5$, and the standardized statistic is $z = -2.13$. What would you expect for the p-value? A. $0.10 <$ p-value B. $0.05 <$ p-value ≤ 0.10 **C. $0.01 <$ p-value ≤ 0.05** D. p-value ≤ 0.01
Relationship between p-value and *z*-score	SBI Exam 1	Suppose you are testing H_0: $\pi = 0.5$ versus H_a: $\pi > 0.5$, and the standardized statistic is $z = 1.79$. What would you expect for the p-value? A. $0.10 <$ p-value **B. $0.05 <$ p-value ≤ 0.10** C. $0.01 <$ p-value ≤ 0.05 D. p-value ≤ 0.01
Definition of p-value	Consensus Practice Exam 4	A research article gives a p-value of 0.001 when reporting the results of a hypothesis test. Which definition of the p-value is most accurate? A. The probability that the observed outcome will occur again. **B. The probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.** C. The value that an observed outcome must reach in order to be considered significant under the null hypothesis. D. The probability that the null hypothesis is true.
Definition of p-value	Consensus Exam 4 SBI Exam 1	The p-value of a test of significance is _____ . A. the probability that the null hypothesis is true B. the probability, assuming the alternative hypothesis is true, that we would get a result at least as extreme as the one that was actually observed **C. the probability, assuming the null hypothesis is true, that we would get a result at least as extreme as the one that was actually observed** D. the probability that the alternative hypothesis is true

Hypotheses are statements about parameters	Consensus Practice Exam 4	A research hypothesis is always expressed in terms of the _____ because we are interested in making statements about the _____ based on a _____ statistic. A. sample statistic; population; sample B. population statistic; population; parameter **C. population parameter; population; sample** D. population parameter; sample; population
Hypotheses are statements about parameters	Consensus Practice Exam 4	The null hypothesis H_0 is the statement of _____ and always has a _____ sign. The alternative hypothesis H_a is the _____ hypothesis. It is a statement about the value of a _____ that we intend to test. A. no consequence; > ; research; sample **B. no change; = ; research; parameter** C. change; = ; no change; parameter D. no change; = ; research; sample
Hypotheses are statements about parameters	Consensus Exam 4	Which of the following is **not** true about the alternative hypothesis? A. It is sometimes called the research hypothesis. **B. It is assumed to be true.** C. Like the null hypothesis, it is always a statement about a population proportion. D. It is usually a statement that the researcher hopes to demonstrate is true.

Interpretation of p-value (in context)	Consensus Practice Exam 4	A janitor at a large office building believes that his supply of lightbulbs has a defect rate that is higher than the defect rate stated by the manufacturer. The janitor's null hypothesis is that the supply of lightbulbs has a manufacturer's defect rate of $p = 0.09$. He performs a test at a significance level of 0.01. The null and alternative hypotheses are as follows: H_0: $p = 0.09$ and H_a: $p > 0.09$. Suppose the janitor tests 300 lightbulbs and finds that 33 bulbs are defective. The janitor calculates a p-value for the hypothesis test of approximately 0.113. Choose the correct interpretation for the p-value. A. The p-value tells us that the probability of concluding that the defect rate is equal to 0.09, when in fact it is greater than 0.09, is approximately 0.113. **B. The p-value tells us that the if the defect rate is 0.09, then the probability that the janitor will have 33 or more defective lightbulbs out of 300 is approximately 0.113.** C. The p-value tells us that the true population rate of defective lightbulbs is approximately 0.113. D. The p-value tells us that the result is significantly higher than the advertised value using a significance level of 0.05.

Interpretation of p-value (in context)	Consensus Practice Exam 4	A researcher is wondering whether the smoking habits of young adults in a certain city in the U.S. are the same as those of the general population of young adults in the U.S. A recent study stated that the proportion of young adults that smoke at least twice a week is 0.16. The researcher collected data from a random sample of 75 young adults in the city of interest. The researcher completes a hypothesis test with a resulting p-value of 0.076. Choose the best statement to interpret the results. A. The p-value for a two-sided test is divided by 2, resulting in a value less than 0.05. The hypothesis that the city of interest has a different proportion of smokers than the general public is plausible. B. The p-value for a two-sided test is multiplied by 2, resulting in a value greater than 0.05. The data suggest that the city of interest has the same proportion of smokers as the general public. **C. The p-value is greater than 0.05, which means it's plausible that the city of interest has the same proportion of smokers as the general public.** D. The p-value is greater than 0.05, which means the data suggest that the city of interest has a different proportion of smokers than the general public.

Interpretation of p-value (in context)	Consensus Practice Exam 4	A medical researcher conducts a hypothesis test to test the claim that U.S. adult males have gained weight over the past 15 years. Assume that all the conditions for proceeding with the hypothesis test have been met. The calculated test statistic is approximately 2.10 with an associated p-value of approximately 0.0179. Choose the conclusion that provides the best interpretation for the p-value at a significance level of 0.05. A. If the null hypothesis is true, then the probability of getting a test statistic that is as extreme or more extreme than the calculated test statistic of 2.10 is 0.0179. This result is **not** surprising and could easily happen by chance. **B. If the null hypothesis is true, then the probability of getting a test statistic that is as extreme or more extreme than the calculated test statistic of 2.10 is 0.0179. This result is surprising and could not easily happen by chance.** C. The p-value is extreme. Therefore, the hypothesis test proves that the null hypothesis is true. D. If the null hypothesis is true, then the probability of getting a test statistic that is as extreme or more extreme than the calculated test statistic of 2.10 is 0.0179. The result should be doubled for a two-sided test. This result is **not** surprising and could easily happen by chance.
Interpretation of p-value (in context)	Consensus Practice Exam 4	A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a p-value of 0.17. Which of the following is a reasonable interpretation of her results? A. This proves that her experimental treatment has no effect on memory. **B. There could be a treatment effect, but the sample size was too small to detect it.** C. She should reject the null hypothesis. D. There is evidence of a small effect on memory of her experimental treatment.

Interpretation of p-value (in context)	Consensus Exam 4	A quality control manager thinks that there is a higher defective rate on the production line than the advertised value of $p = 0.025$. She performs a hypothesis test with a significance level of 0.05. Symbolically, the null and alternative hypotheses are as follows: H_0: $p = 0.025$ and H_a: $p > 0.025$. She calculates a p-value for the hypothesis test of defective light bulbs to be approximately 0.067. Which of the following is the correct interpretation of the p-value? **A. The p-value tells us that if the defect rate is 0.025, then the probability that she would observe the percentage she actually observed or higher is 0.067. At a significance level of 0.05, this would not be a surprising outcome.** B. The p-value tells us that the probability of concluding that the defect rate is equal to 0.025, when in fact it is greater than 0.025, is approximately 0.067. C. The p-value tells us that the true population rate of defective light bulbs is approximately 0.067. D. The p-value tells us that the result is significantly higher than the advertised value using a significance level of 0.05.
Interpretation of p-value (in context)	SBI Practice Exam 1	Consider the Doris and Buzz study, where the researcher was investigating whether dolphins can communicate. If the results of a simulation return a p-value of 0.15, what should we conclude? **A. The sample results could have happened by chance. The sample evidence does not support the conjecture that dolphins can communicate.** B. The sample results are unlikely under the "by chance" model. The sample evidence suggests that dolphins can communicate. C. The alternative hypothesis is plausible. Buzz could have just been guessing at which button to press. D. The null hypothesis is plausible. Buzz seems to understand communications from Doris.

Interpretation of p-value (in context)	SBI Exam 1	Consider the Doris and Buzz study, where the researcher was investigating whether dolphins can communicate. If the results of a simulation return a p-value of 0.03, what should we conclude? A. The sample results could have happened by chance. The sample evidence does not support the conjecture that dolphins can communicate. **B. The sample results are unlikely under the "by chance" model. The sample evidence suggests that dolphins can communicate.** C. The alternative hypothesis is plausible. Buzz could have just been guessing at which button to press. D. The null hypothesis is plausible. Buzz seems to understand communications from Doris.
Interpretation of p-value (no context)	Consensus Exam 4 SBI Exam 1	When we get a p-value that is very small, we may conclude that _____ . A. the null hypothesis is plausible B. the alternative hypothesis has been proven to be false **C. there is strong evidence against the null hypothesis** D. there is strong evidence against the alternative hypothesis
Interpretation of p-value (no context)	SBI Quiz #2	When we get a p-value that is very small, we may conclude that… A. the null hypothesis has been proven to be true. **B. there is strong evidence for the alternative hypothesis.** C. the null hypothesis is plausible. D. the alternative hypothesis has been proven to be false.