COMMUNITY COLLEGE INSTRUCTORS' AND STUDENTS' EXPLORATION OF SIMULATION-BASED STATISTICAL INFERENCE

A Capstone Project

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

---

by

Irina Timchenko, B.S., M.S.,

May, 2017

Curriculum, Instruction, and Special Education
Curry School of Education
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE CAPSTONE PROJECT

This capstone project, "Community College Instructors' and Students' Exploration of Simulation-based Statistical Inference," has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Education.

_____
Name of Chair (Joe Garofalo)

_____
Committee Member Name (Susan Mintz)

_____
Committee Member Name (Robert Berry)

_____Date

ACKNOWLEDGEMENT

I could not have completed this work without the help and support of my loving family, supportive colleagues, talented and caring Curry faculty, encouraging friends, and motivating students.

I am especially grateful to my mother Marina for her support, patience, and help; to my advisor Joe for his guidance, encouragement, and sense of humor; to my daughter Anu for giving me energy and strength; and to my friend Wendi, for her motivation and feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Statistics education is widely considered a necessary area of study due to the prevalence of statistical information in today's world. Introductory statistics courses are especially important because they provide a foundation for conducting and interpreting quantitative research. Unfortunately, student encounter difficulty grasping statistical concepts and especially struggle with statistical inference. Students too often memorize steps without fully comprehending the reasons behind the procedures and are unable to retain and transfer information. To solve this problem, many statistical educators recommend using computer simulations to help students visualize statistical concepts and to emphasize the logic of inference. However, the research on the effectiveness of simulation-based inferential methods is sparse. In addition, there is a disagreement between statistics educators about using simulations for instruction. While some educators recommend both formula- and simulation-based instructional methods, others advocate implementing only simulation-based approach.

The purpose of this study was to investigate students' thinking, reasoning, and views on simulation-based methods of statistical inference. In addition, the study examined instructors' attitudes toward teaching hypothesis testing through computer simulations.

The study participants consisted of four statistics instructors and six students at a community college in Virginia (CCCC). The instructors were interviewed about their experience with teaching statistics and their views on their students' understanding of and

difficulties with statistical inference. They were shown examples on how to use simulations designed by Rossman and Chance (2009) to make inference about a population proportion and to compare two population means. Afterwards, the instructors shared their reactions toward each simulation and provided recommendations on implementing simulation-based inference (SBI) in introductory statistics courses.

Four of the interviewed students completed an introductory statistics course at CCCC, while two students had no previous statistical background. All six students were presented with three tasks and were guided through the investigation process. After that, students were asked to share their views on simulations.

The instructor and student cases were examined in a cross-matrix to develop multi-case assertions. Results from the instructor cross-case analysis included:

1) Instructors believe that their students struggle most with statistical inference because of their inability to understand statistical concepts and interpret results as well as their difficulty with selecting an appropriate test for a problem.
2) All instructors expressed mostly positive views of SBI, citing various benefits of the simulations.
3) Most instructors regard the binomial simulation as a better statistical approach than its corresponding formal test. They view the randomization test as either comparable to its corresponding formal 2-sample *t*-test, or have no opinion on this matter.

4) Instructors provided different recommendation for using SBI in the classroom ranging from using simulations as a substitute for traditional methods to not using them at all.

Results from the student cross-case analysis included:

1) Most students who took a statistics course reported having difficulty with hypothesis testing and retained minimal and mostly factual knowledge shortly after taking the course.
2) Most students were able to understand screen representations but struggled with comparing results of the experiment with the empirical distribution and with making conclusions.
3) After using the simulations all students who took an introductory statistics course made some connections between SBI and formal methods and understood better the concepts learned in their statistics classes.
4) All students expressed mostly positive views on SBI, citing various benefits of the simulations.
5) Most students reported that formal and informal methods have their own strength and weaknesses. For this reason, five out of six students expressed preference for a blended learning approach.

Based on the study results, I propose the following recommendations:

1) A combination of formal and informal inferential methods should be used in an introductory statistics course.
2) Simulation-based approach should be introduced before formal methods.
3) Instructors should be offered a professional development opportunity for understanding and implementing SBI.
4) Continue to explore student understanding and difficulties with SBI.

1. PROBLEM OF PRACTICE

Introduction

With the growing presence of quantitative information and statistical claims in today's world, people need to be able to understand and intelligently evaluate data-based arguments. The field of statistics offers methods for using data to describe a phenomenon and to make informed decisions. In order to produce statistically literate individuals who can understand and also apply statistical methods, introductory statistics courses have been widely offered in high schools as well as universities. According to the Conference Board of Mathematical Sciences' (CBMS) 2010 national survey of undergraduate education, overall enrollment in probability and statistics courses from 2005 to 2010 has increased from 260,000 to 370,000 students in four-year colleges and universities and increased from 117,000 to 137,000 in two-year colleges students (CBMS, 2010).

Due to the prevalence of numerical information and increased number of students taking statistics courses, statistics education is a necessary and important area of study. In recent years, statistical researchers and educators have focused their attention on developing students' statistical reasoning and conceptual understanding of statistical ideas (e.g. Garfield & Ben-Zvi, 2007; Nikiforidou, Lekka, & Pange, 2010). They suggested improving student understanding of statistical concepts by creating student-centered activities, making students aware of their errors, using appropriate technological tools, and providing consistent and helpful feedback (Garfield & Ben-Zvi, 2007).

National organizations, such as The National Council of Teachers of Mathematics (NCTM) and The American Statistical Association (ASA), also recognize the need for quality statistics education. NCTM (2000) has included data analysis and probability standards as early as in elementary grades. The Board of Directors of the ASA endorsed Guidelines for Assessment and Instruction in Statistics Education (GAISE) for the teaching of an entry-level high-school and college statistics courses. According to the GAISE college report (ASA, 2005), "the desired result of all introductory statistics courses is to produce statistically educated students, which means that students should develop statistical literacy and the ability to think statistically" (p.11). In order to achieve this goal, the report recommends utilizing technology, using real data, stressing conceptual understanding of statistical ideas, and fostering active learning in the classroom.

Clearly, researchers and national organizations emphasize developing students' statistical thinking and reasoning rather than mere memorization of formulas and procedures.

Statement of Problem

In spite of the growing importance of statistics and recommendations from national organizations, students find the subject challenging and counterintuitive. They often reason incorrectly about statistical ideas and are unable to understand and remember basic statistical concepts (e.g. Garfield & Ben-Zvi, 2008). Students especially struggle with inferential reasoning as it requires understanding multiple concepts

including descriptive statistics, basic probability rules, and sampling distributions (Castro Sotos, Vanhoof, Noorgate, & Onghena, 2009; Castro Sotos, Vanhoof, Noortgate, & Onghena, 2007, Garfield & Ben-Zvi, 2008). In addition, learners focus on numbers, computations, and formulas, which limits their ability to connect different concepts and make sense of statistical information (Garfield & Ben-Zvi, 2008). When students are first taught the formal method of hypothesis testing, they are introduced to many new terms (null and alternative hypotheses, critical value, critical region, Type I, Type II errors, etc). Learning new terminology places a high cognitive demand on students and makes it difficult for them to understand the main idea of inference. Therefore, students too often memorize steps without fully comprehending the reasons behind the procedures and are unable to retain and transfer information.

Attempts to Solve the Problem

Many researchers suggest reforming not only teaching methods, but also the content of introductory college-level statistics courses in order to improve students' statistical reasoning and thinking (Carver, 2011; Cobb, 2007; Garfield, delMas, & Zieffler, 2012). They recommend eliminating elementary probability, manual computations, and reading tables to allow time for in-depth statistical investigations as shown in Table 1. A typical introductory statistics course consists of descriptive and inferential statistics as well as elementary probability concepts. While descriptive statistics is used to summarize and describe data, inferential statistics involves using a

sample to either estimate a population parameter (confidence intervals) or to evaluate a research hypothesis (hypothesis testing).

Table 1

Carver's Proposed Curricular Changes (Carver, 2011, p.3)

OUT: Topics We Can Afford to Drop	**In: Formerly "Advanced" or Neglected Topics**
• Most manual computations using small datasets • All software-based computation using artificial data • Defining histogram bins • Most elementary probability • Reading tables of *t*, *z*, *F*, etc.	• Data preparation, missing data, and data management • Writing about statistical investigations • Resampling and Permutation tests • Analysis of Probabilistic non-SRS data using sampling weights • Nonparametric methods • Non-linear models • Multivariate models

Rather than teaching descriptive and inferential statistics separately as is done in a traditional introductory statistics curriculum, some researchers (Carver, 2011; Cobb, 2007; Garfield, Zieffler, & Ben-Zvi, 2015) propose emphasizing the logic of inference throughout the course, which will make the curriculum more coherent and less compartmentalized, resulting in students making connections between statistical concepts. Researchers suggest using "informal" methods of inference in order to bridge the gap between descriptive and inferential statistics. While formal inferential methods require computations, the use of formulas, and reference to statistical tables, informal approaches rely only on students' intuitive reasoning and a "big picture" of hypothesis

testing. Students make judgements, claims, or predictions about a population based on samples, without formal statistical procedures and methods. An informal approach offers simplicity and access to understanding inference. According to Pfannkuch et al. (2011), informal methods of inference "will decrease cognitive load by reducing the number of concepts that need to be activated simultaneously" (p. 911).

Many educators recommend using computer simulations for teaching inference in introductory statistics courses. While some researchers suggest using informal simulation-based inference as an introduction to the formal methods of constructing confidence intervals and testing hypothesis (Garfield, delMas, & Zieffler, 2012), others propose completely replacing formal methods with the informal inferential approach (Carver, 2011; Cobb, 2007). In either case, it is important to examine student reasoning about informal inferential ideas and whether this method of teaching improves learners' understanding of statistical concepts.

Purpose of Study

Although students' difficulty with inference is widespread throughout the nation, the problem is also very profound at a local community college where I teach mathematics and statistics. Therefore, it is important to explore the alternative ways for teaching inference. The purpose of this proposed capstone project is to investigate student thinking and reasoning about simulation-based methods of inference at the college. The study will also examine the teachers' views of their students' current understanding of inferential statistics and their reaction to informal methods of inference. The research

questions and the conceptual framework are presented in the next chapter after the literature review.

Definition of Terms

The following is a list of definitions of key terms used in the context of this study.

- Statistical Inference – a process that aims at learning characteristics of the population from a sample. Hypothesis testing is one method of statistical inference
- Hypothesis Testing – a procedure that uses a sample data to draw conclusions about one or more populations
- Computer Simulation – a computer program that provides a model of hypothesized or natural phenomena that allows users to explore the implications of manipulating parameters within the program.
- Applet – a web-based computer application. The term is often used synonymously with simulation.
- SBI- simulation-based inference.
- GAISE – the Guidelines for Assessment and Instruction in Statistics Education. A framework for statistics education for pre-K–12 and college-level students. The reports are published by ASA.
- ASA – The American Statistical Association. It is the world's largest community of statisticians that supports the excellence in the development, and dissemination of statistical science through meetings, publications, and membership services (http://www.amstat.org/about/index.cfm)

2. LITERATUFRE REVIEW

Hypothesis Testing Theories

The method of hypothesis testing, also referred to as significance testing, uses sample data to draw conclusions about one or more populations. Although statistics textbooks typically present a single model of statistical inference, there are several views of hypothesis testing, which differ philosophically and methodologically (Hubbard & Bayarri, 2003). In Ronald Fisher's model the researcher sets up a null hypothesis, which states that a sample comes from a hypothetical population with a known sampling distribution. The researcher then proceeds to calculate the test statistic that measures the difference between what is observed and the null hypothesis. The test statistic is then converted to a probability, called a *p-value.* This probability measures the likelihood of obtaining the observed, or more extreme difference, given the truth of the null hypothesis. If this probability is low (typically less than 5%), then the null hypothesis is rejected or disproved. A low probability of obtaining the sample data given the null hypothesis is true indicates that either a very rare event occurred or the null hypothesis is false (Fisher, 1959). As Hubbard & Bayarri (2003) explain, "A *p* value for Fisher represented an objective way for researchers to assess the plausibility of the null hypothesis" (p. 172).

Jerzy Neyman and Egon Pearson intended to improve Fisher's theory and put forward a different model known as Neyman-Pearson approach (Hubbard & Bayarri, 2003). According to this model, there are two competing hypotheses, the null (H_0) and the alternative hypothesis (H_A). It should be noted that the alternative hypothesis is not

explicitly stated in the Fisher's approach and was a source of disagreement between the two perspectives.

Another difference is that the Neyman-Pearson hypothesis testing framework specifies errors researchers are willing to accept. There are two errors possible when making a decision about the rejection of the null hypothesis. The first error (Type I error) is false rejection of the null hypothesis, and the second error (Type II error) deals with false acceptance of the null hypothesis. The probabilities of Type I and Type II errors are denoted by α and β respectively. Additionally, the Neyman-Pearson approach introduces the concept of statistical power $(1 - \beta)$, which gives the probability of correctly rejecting the null hypothesis.

Rejecting the true null hypothesis is considered to be a more serious error. For this reason, researchers specify a significance level before testing a hypothesis and typically set it at 5%. The selected probability of type I error determines the rejection region for the null hypothesis. The rejection region specifies the values of the test statistic which lead to rejection of the null hypothesis. The more the data deviates from the null hypothesis, the higher the test statistic, and the more likely to refute the null hypothesis. The critical value is a point that separates the rejection region from the non-rejection region.

Biau, Jolles, and Porcher (2010) summarize the differences between Fisherian and Neyman-Pearson's approaches as presented in Table 2.

Table 2

Comparison of Fisher's and Neyman-Pearson's Approaches (Biau, et al., 2010. p. 887).

Fisher's p value	Hypothesis testing
Ronald Fisher	Jerzy Neyman and Egon Pearson
Significance test	Hypothesis test
p Value	α
The p value is a measure of the evidence against the null hypothesis	α and β levels provide rules to limit the proportion of errors
Computed a posteriori from the data observed	Determined a priori at some specified level
Applies to any single experiment	Applies in the long run through the repetition of experiments
Subjective decision	Objective behavior
Evidential, ie, based on the evidence observed	Nonevidential, ie, based on a rule of behavior

Hubbard & Bayarri (2003) explain that Fisher's approach is inductive in nature as it moves from specific to general. Based on the extremity of sample results, the researcher makes inferences about characteristics of a population. On the other hand, Neyman-Pearson's approach argues from the general to the particular because it establishes rules "for choosing between two alternative courses of action, accepting or rejecting the null hypothesis" (p. 273).

Another major difference exists between the concepts of Fisher's *p-value* and Neyman-Pearson's α. While α is specified prior to the collection of data, *p-value* is a

random variable over the interval [0,1] as it is calculated based on the sample obtained from the experiment. Moreover, a *p-value* applies to a single experiment and refers to the probability of obtaining data this extreme (or more extreme) under the null hypothesis. On the other hand, Neyman and Pearson's α deals with a long-run frequency of experiments and does not provide evidence by itself for rejection or acceptance of a particular hypothesis. A 5% type I error rate indicates that "of all the tests being carried out around the world, at most 5% of them result in a false rejection of the null" (Hubbard & Bayarri, 2003, p. 176). The Neyman-Pearson framework is concerned with minimizing errors, which only applies to long-run repeated sampling situations, not to individual experiments.

Both *p-value* and Type I error rate α are both tail probabilities, and therefore are often confused by students and researchers. Moreover, since *p-value* is sometimes referred to as an observed error rate, it is often mixed up with α as they both are viewed as errors (Hubbard & Bayarri, 2003).

Despite the differences in the two theories of hypothesis testing, contemporary approaches blend the methods. This often leads to misuse of the original approaches not only by students, but also novice researchers and statisticians (Biau et al., 2009; Hubbard & Bayarri, 2003; Shelskin, 2007). Textbooks typically present Fishers' idea of disproving the null hypothesis, but use Neyman-Pearson concepts of alternative hypothesis, Type II errors, and the power of statistical tests. Most textbooks fail to acknowledge the embedded model of the hypothesis testing theory, creating confusion not only because of

a large amount of new concepts, symbols, and definitions, but also due to the conflicting philosophies (Hubbard & Bayarri, 2003).

Understanding of the theory of hypothesis testing is critical for planning, conducting, interpreting, and reporting scientific experiments (Biau et al., 2009). Unfortunately, many students struggle to comprehend the concepts and logic of hypothesis testing. The section below reviews student difficulties in this area.

Student difficulties with Hypothesis Testing

While students typically are able to perform a formal procedure of hypothesis testing, they are rarely able to grasp its meaning (Castro Sotos, Vanhoof, Noortgate, and Onghena, 2009; delMas, Garfield, Ooms, & Chance, 2007; Haller & Krauss, 2002). Student errors have been documented during all stages of hypothesis testing including misconceptions concerning the definition of hypotheses, the significance level, the interpretation of a *p*-value, and the logic of the overall hypothesis test. These difficulties are discussed below.

Castro Sotos et al. (2009) investigated students' misconceptions regarding the definition of hypothesis test, *p-value*, and significance level. They administered a questionnaire to 144 (95 females, 48 males) university undergraduates majoring in mathematics (21%), medicine (52%), and business (23%) who recently completed an introductory statistics course. The questionnaire consisted of five multiple-choice conceptual questions about hypothesis testing. The researchers found that students lacked understanding of what a hypothesis test means. For example, 21% believed that it is a

mathematical proof of the null hypothesis, and 19% thought it is a probabilistic proof by contradiction. Similarly, students had difficulty interpreting *p-value* and significance level. Twenty-one percent of the students identified *p-value* as the probability of the null hypothesis and 16% interpreted it as the probability of incorrectly rejecting the null hypothesis. As far as significance level, 17% of students interpreted it as the probability of the null hypothesis and 17% indicated that 1- α is the probability of rejecting the null hypothesis.

Unfortunately, many students do not gain an understanding of statistical ideas after completing an introductory college-level course. In order to measure students' conceptual understanding of essential statistical ideas, delMas et al. (2007) developed the Comprehensive Assessment of Outcomes in Statistics (CAOS) test consisting of 40 multiple choice items. With Cronbach's alpha coefficient of 0.82, the assessment exceeded suggested internal consistency levels. CAOS was administered to 763 introductory statistics students in 20 different two-year and four-year institutions from 14 different states before and after completing the course.

The study showed that students had a very low performance on both pre- and post- tests on items regarding hypothesis testing. About half of the students incorrectly interpreted results of a significance test when the null hypothesis was rejected on the post-test. Moreover, even though 58.6% of the students recognized a correct interpretation of a *p-value*, the majority of these students also chose an incorrect interpretation, failing to recognize a contradiction between the two interpretations.

Students also demonstrated difficulty with concepts of sampling distribution and sampling variability, which are fundamental for understanding hypothesis testing (Garfield, Le, Zieffler, & Ben-Zvi, 2014).

Another common misconception deals with the interpretation of the type I error as the probability of the null hypothesis being true given that it has been rejected. The confusion results from switching the conditional statement as the correct interpretation is the probability of rejecting the null hypotheses given that it is true. In a study by Vallecillos (2002), 53% of 436 college students agreed with the following incorrect statement: “A significance level of 5% means that, on average, 5 times out of every 100 times we reject the null hypothesis, we will be wrong.”

With regards to the mistakes about role of the hypotheses, many students mistakenly believe that (a) Null hypothesis is the hypothesis to be proved, (b) Null hypothesis refers to either the population or the sample, and (c) Null hypothesis refers to only one population or only one parameter (Vallecillos, 1999).

In summary, students hold numerous and deep misconceptions in statistical inference. Table 3 summarizes and classifies student errors based on reviewed studies. It is important to understand the source of these errors, which is addressed in the next section.

Table 3

Summary of Student Misconceptions about Hypothesis Test

Nature of Hypothesis Test • Mathematical proof of the null hypothesis • Probabilistic proof by contradiction • A proof of the probability or improbability of H_0
Hypotheses • Null hypothesis refers to either the population or the sample • Null hypothesis refers to only one population or only one parameter
Significance Level • α is the probability of the null hypothesis • 1- α is the probability of the null hypothesis • α is the probability of the null hypothesis being true given that it has been rejected • Confusing the significance level with the *p-value*
P-value • If *p-value* is smaller than alpha, then a definitive statement about a hypothesis can be made • *p-value* is the probability of the null hypothesis being true • *p-value* is the probability of incorrectly rejecting the null hypothesis • *p-value* is the probability of obtaining the same or more extreme results • A small *p-value* means the results have significance (statistical and practical significance are not distinguished)

The Source of Student Misconceptions

Misinterpretation of inferential concepts is widespread among students. Possible reasons for student misconceptions are (1) a complex nature of hypothesis testing, (2) teachers' misunderstanding of inference, and (3) misinterpretation of statistical concepts in textbooks. Each of these sources is discussed below.

Complex Nature of Hypothesis Testing

One reason for student misunderstandings could be the fact that hypothesis testing involves not only understanding, but also relating many abstract concepts. In addition, as mentioned earlier, hypothesis testing is presented as a hybrid between Fisher's and Nyman-Pearson's approaches, which could result in a confusion (Castro Sotos, Vanhoof, Noortgate, and Onghena, 2007). Furthermore, inconsistent and incorrect terminology, such as referring to *p-value* as an error rate, could be the cause of misunderstanding (Hubbard & Bayarri, 2003).

Castro Sotos et al. (2007) provided possible explanations for some of the above stated errors. For example, students consider hypothesis testing as a mathematical proof because they believe that as any mathematical proof, the test results are also deterministic. As a result, they believe that the null hypothesis is proven to be either true or false. Some students view hypothesis testing as proof by contradiction because of their similarity in structure or reasoning. The proof by contradiction is based on the logical *modus tollens* method, which states that if *P* implies *Q*, then the contradictory of *Q* implies the contradictory of *P*. The authors give the following example. If it is raining

there are clouds. Therefore, if there are no clouds, it is not raining. Students apply similar logic to hypothesis testing. They reason that if H_0 is true, then there is a high probability that the *p-value* is large. Therefore, if *p-value* is small, then H_0 is improbable. However, the authors explain that this reasoning is incorrect because "a low probability event does not make the premise from which it is drawn improbable" (p. 106). The logic only works in case of deterministic events (when probabilities are zero and one).

According to Thompson, Saldanha, and Liu (2004), students struggle with inference because they have to attend to two aspects of a sample. First, they have to focus on one sample and its statistical summary and second, they need to understand variability of a summary measure among collection of samples. Unfortunately, students find the second task challenging and are unable to attend to variability between samples (Garfield, Zieffler and Ben-Zvi, 2015; Lee, Angotti, and Tarr, 2010).

Teacher Difficulties with Hypothesis Testing

The concepts of hypothesis testing are challenging not only for students, but also for researchers and instructors. Haller and Krauss (2002) presented six wrong statements about hypothesis testing to a group of statistics and methodology instructors, scientists, and psychology students in six German universities. They found that 80% of methodology instructors, 90% of scientists, and 100% of students chose at least one of the false statements as correct. The most popular incorrect statement was "you know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision" (p.5). This idea appears similar to the definition of Type I error (the probability

of rejecting the true null hypothesis). However, having actually rejected the null, this decision would be wrong if and only if H_0 were true. Thus, the given statement is equivalent to the knowledge of the probability of the null hypothesis being true, which we cannot gather from performing a hypothesis test.

Liu and Thompson (2005) examined understanding of hypothesis testing of eight high-school mathematics teachers who had at least a BA in mathematics or mathematics education and statistics teaching experience. Teacher interviews revealed that most of the teachers did not understand the logic of hypothesis testing as their knowledge of probability and statistical inference was highly compartmentalized. The teachers struggled with the concept of statistical "unusualness" within hypothesis testing and had difficulty deciding for what types of questions hypothesis testing methods were appropriate. Most of the teachers also demonstrated "commitment" to the null hypothesis and did not want to reject it, even with a very small *p*-value. As authors explained, the teachers did not want to reject the null hypothesis on the basis of one sample because "any rare sample could still occur theoretically" (p. 4).

Misrepresentation of Statistical Concepts in Textbooks

Another explanation for student difficulties could be misrepresentation of statistical concepts in textbooks. Glinger, Leech, and Morgan (2002) reviewed a diverse set of twelve textbooks used in graduate-level research and statistics courses in an education department. At least two researchers reviewed each book and produced high interrater reliability ratings. The authors reported that only 17% of textbooks addressed

the controversy between Fisher's and Neyman-Pearson's methods of inference. In addition, 33% of the textbooks failed to emphasize that the size of the *p-value* does not indicate the strength of treatment, which is a common student error. While some of the texts mentioned effect size, most did not. Finally, although a majority of textbooks (84%) addressed the difference between statistical and practical significance, it was often unclear how to assess practical significance of a result.

Without a doubt, student understanding of hypothesis testing needs to be improved because drawing inferences from data and evaluating results of research studies is an important skill for all adults (Garfield & Ben-Zvi, 2008). The following section will present several suggestions on enhancing students' understanding of this topic.

Improving Student Understanding of Inference

There are several recommendations for teaching inference to make it more accessible for students. For instance, Garfield and Ben-Zvi (2007) recommend shifting from teaching computations to enhancing student's conceptual understanding of statistical ideas. They suggest implementing student-centered activities, making students aware of their errors, using appropriate technological tools, and providing consistent and helpful feedback.

Rossman and Chance (1999) claim that in an introductory statistics course inference is being taught as an isolated subject without any connection to the exploratory data analysis and suggest the following "Top Ten" list of recommendations for teaching hypothesis testing:

1) Insist on complete presentation and interpretation of results in the context of the 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

The authors argue that computer simulations are more effective than formal probability for introducing students to the concepts of sampling distribution and hypothesis testing. They also advocate using the *p-value* method when making decisions and emphasizing reasoning process underlying inference formulas.

Garfield and Ben-Zvi (2008) believe that by integrating the foundations of statistical inference throughout the course, students will be "less confused by the formal ideas, procedures, and language when they finally reach the formal study of this topic" (p. 264). In order to help students understand the process of testing hypothesis, the authors suggest using the logic of argumentation. They view hypothesis testing as a method for supporting an argument. The argument is about claims (hypotheses), which can be either true or false. While the researchers cannot prove whether the hypothesis is true or false, they can gather evidence to support the argument. Garfield and Ben-Zvi provide the following four "building blocks" to make a convincing argument:

1. A clear claim we are making (and a counterclaim that includes all other possibilities).
2. Data to support our argument.
3. Evidence that the data are accurate and reliable, not misleading.
4. A good line of reasoning that connects our data to our argument. (p. 271)

The authors explain that just like in real life whether we win or lose an argument depends on the strength of the evidence for our claim, in hypothesis testing we also use data as an evidence for or against a hypothesis. The farther is our data from the claim we are arguing against, the more evidence we have to reject that claim.

The recommendations presented above are closely related to the idea of Informal Statistical Inference (ISI), which has been an interest for many researchers as well as teachers of statistics in recent years, and is discussed in the following section.

Informal Inference

Many researchers suggest using informal inference as a bridge between descriptive statistics and formal inferential methods (Makar and Rubin, 2009; Zieffler, Garfield, delMas, and Reading (2008). There are various definitions of Informal Inference. Makar and Rubin (2009) define informal inference as "the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected" (p. 83). According to Pfannkuch (2006), ISI is "the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions for data" (p. 149). Finally, Zieffler et al. (2008) view informal inference as "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (p.44). These

definitions imply that ISI involves making judgements, claims, or predictions about a population based on samples, without using formal statistical procedures and methods.

Makar and Rubin (2009) suggest introducing students to ISI at an early age long before they are taught formal inferential techniques. The authors identify three key features that form statistical inference: (1) Generalization beyond describing the given data, (2) Use of data as evidence to support these generalizations, and (3) Employment of probabilistic (non-deterministic) language that supports some level of uncertainly about the conclusions drawn.

In the first principle the authors argue that inference should extend beyond descriptive statistics and should make a claim about a wider universe. While we often use descriptive statistics as a means to describe the available data, the main purpose of statistics is looking beyond the data to make estimates and predictions. The second principle stresses the need for using strong forms of evidence rather than mere anecdotes or beliefs. Finally, the third feature emphasizes that our generalizations are not definite since we only consider a sample drawn form a larger population. Makar and Rubin express that the level of uncertainty does not have to be quantified with younger children and can be addressed with the introduction of confidence interval or *p-value*.

Another relatively new concept in research literature related to ISI is the idea of Informal Inferential Reasoning (IIR). Zieffler et al. (2008) define IIR as "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (p. 44).

Makar, Bakker, and Ben-Zvi (2011) present three key elements that support students' Informal Inferential Reasoning (IIR). These elements are (1) An inquiry-based learning environment, (2) A set of cognitive and social elements, and (3) A conflict between students' expectations and statistical results.

The design of inquiry-based environment incorporates relevant tasks, technological tools, and scaffolds by the teacher and educational materials. The cognitive features include individual and collaborative reasoning processes in identifying beliefs and doubts, resolving conflicts, formulating questions, drawing conclusions, and coming up with explanations. On the other hand, the social elements encompass social and statistical norms, including seeking peer consensus and clarification, asking meaningful questions, and searching for explanations. Finally, students' doubts as they obtain unexpected results create a cognitive conflict, which propels their inquiry and search for explanations.

In order to study students' informal inferential reasoning, Zieffler et al. (2008) suggest to utilize tasks that challenge students to

1. Make judgments, claims, or predictions about a population based on samples, but not using formal statistical procedures and methods (e.g., p-values, *t* tests);
2. Draw on, utilize, and integrate prior knowledge (formal and informal) to the extent that this knowledge is available; and
3. Articulate evidence-based arguments for judgements, claims, and predictions about populations based on samples (p.46).

The first component helps reveal whether the student made reasonable inferences about one or more populations based on one or more samples. The second component helps shed light on how students making inferences integrate informal knowledge, such

as everyday knowledge about the problem context, prior statistical knowledge, real world knowledge and experience, and statistical language. Finally, the third item examines how the student use the evidence to substantiate his/her arguments in making conclusions and drawing inferences and how well the evidence supported the inferences made.

Researchers advocate using informal inferential methods not only with young children, but also in college-level introductory statistics classes (Carver, 2011; Cobb, 2007; Tintle, 2015). These methods are based on data-simulations and are discussed below.

Simulation-Based Informal Inference

Some educators suggest radically different approaches to statistical inference in introductory, university-level statistics courses. They recommend using simulation methods to build informal inferential reasoning without focusing on the details of formulas and computations (Cobb, 2007; Garfield and Ben-Zvi, 2008; Tintle, 2015), and propose completely abandoning formal inference.

Simulation-based inference (SBI) typically refers to binomial simulations, randomization tests, or bootstrapping. Binomial simulations are used to conduct significance test of a single proportion. Randomization (also called re-randomization or permutation) tests are utilized to simulate the re-randomization of subjects to treatment groups in order to evaluate statistical significance of an observed treatment effect (Tintle, 2015). Finally, bootstrapping illustrates sampling variability and determines confidence intervals without relying on a theoretical sampling distribution. Simulation-based

inferential methods produce the distribution of any test statistic specified under a null hypothesis by reshuffling the original data many times. They are not restricted to the mean and can be used for a variety of quantities of interest, such as medians, quartiles, and measures of variation (Cobb, 2007). These methods differ from formal inferential methods in that they do not rely on asymptotic approximations (Budgett, Pfannkuch, Regan, & Wild, 2013).

George Cobb is the driving force behind the movement toward teaching introductory statistics with SBI. Cobb (2007) criticizes existing curriculum for being centered on the normal distribution instead of on the logic of inference, making the subject confusing and incoherent. He argues that before the existence of computers, researchers had to use analytical methods of inference since computation was impossible. However, due to the advancement of instructional technology and the increase in computational power, it is no longer necessary to assume normality of a sampling distribution or to make corrections to allow for non-normality. Instead, he proposes abandoning formal methods of hypothesis testing and basing inference on data simulation following the three R's of inference model: "Randomize, Repeat, and Reject any model that puts your data in its tail" (p. 12). The idea is for the students to generate a random sample and, if applicable, to split data randomly into groups, then randomize data production over and over to see which outcomes are typical, which are not, and, finally, reject the hypothesized model if the original sample is unlikely.

Based on Cobb's (2007) recommendations on teaching statistics, Carver (2011) offers an introductory statistics curriculum that is designed to improve students' statistical thinking. He suggests taking out some topics (e.g. manual computations, elementary probability, reading tables) in order to free up time for in-depth statistical investigations, nonparametric tests, and non-linear and multivariate models. Carver's course design follows Plan-Do-Report (P-D-R) iterative cycle, during which students (1) raise data driven question and plan for data collection, (2) perform statistical analysis using both formal and informal techniques, and (3) report and explain the results.

Other researchers were also inspired by Cobb's ideas. Tintle, Topliff, VanderStoep, Homes, and Swanson (2011) redesigned the introductory statistics course with an emphasis on the logic of inference using a randomization approach to hypothesis testing. The new curriculum introduced an informal approach of inference early in the course and was later connected to formal asymptotic tests.

Similarly, Garfield, delMas, and Zieffler (2012) developed a simulation-based curriculum called Change Agents for Teaching and Learning Statistics (CATALST). From the first day of instruction, students using this curriculum make informal inference by creating a model for a problem, simulating data from that model, and using the developed distribution to make statistical inference.

The next section provides examples of the binomial and randomization methods of inference.

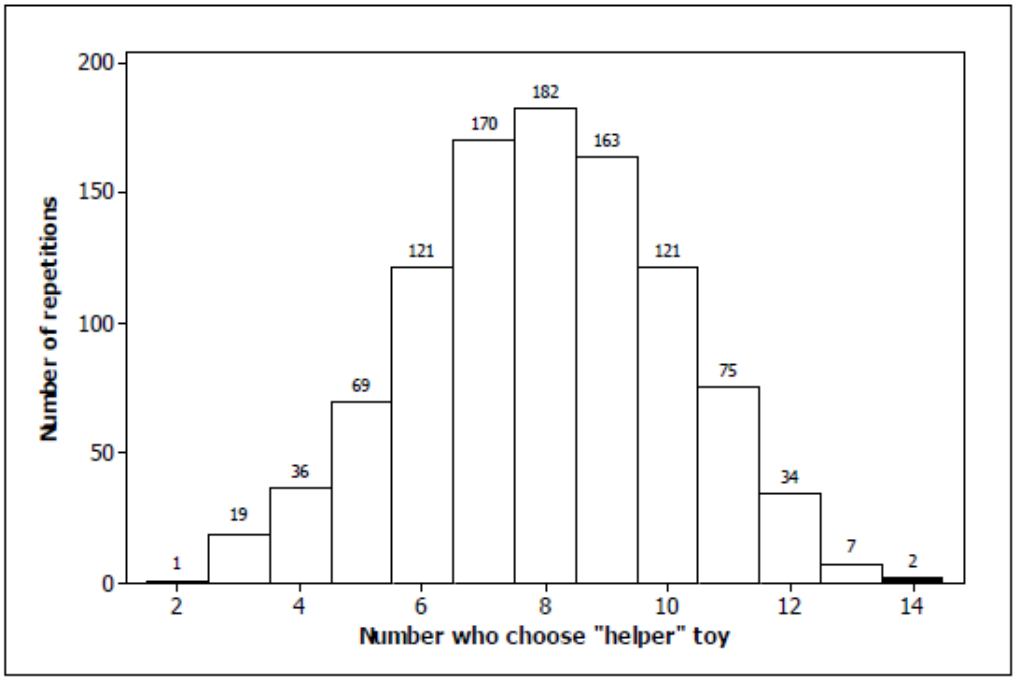
Examples of Simulation-Based Inference

Binomial simulation. A commonly used example of binomial simulation is a "Toy Preference" problem borrowed from Hamlin, Wynn, and Bloom's (2007) study, in which 10-month-old infants were shown a character trying to climb a hill. Then, children were shown another two scenarios. One, in which another character ("helper") pushes up the climber, and second, in which the climber was pushed back down the hill by a different character ("hinderer"). After that, sixteen infants were offered either "helper" or "hinderer" character to play with. Fourteen infants chose the helper toy.

Rossman (2008) demonstrated how to investigate whether infants prefer the helper to the hinderer using a computer simulation. He explained that while more than half of the infants chose the helper toy, this selection could have occurred by chance. Assuming that infants have no genuine preference (null model), Rossman suggested to model the infants' selection as flips of a fair coin, but on a computer. He simulated 16 fair coin flips 1000 times and recorded the number of heads (which corresponds the number of infants choosing the helper toy) occurring in each simulation as shown in Figure 1 below. The histogram reveals that out of 1000 groups of 16 infants only two samples provide the same results as the one in the study (14 helper toy selections), assuming no genuine preference. Given that it is very unlikely to find 14 or more choosing the helper toy, there is strong evidence to conclude that infants genuinely tend to prefer the helper toy over the hinderer toy.

Figure 1

Binomial simulation for helper toy study (Rossman, 2007. p. 8).



Randomization Method. Pfannkuch et al. (2011) provide an example that compares diastolic blood pressure of individuals following fish oil and regular oil diets based on Knapp and Fitzgerald's study (1989). Fourteen males with high blood pressure were randomly assigned to either a fish oil (treatment) or a regular oil (control) diet, each lasting four weeks. After that, the diastolic blood pressure reduction level was calculated for each individual in millimeters of mercury. The average reduction level of males in the treatment group was 7.71 mmHg higher than the average reduction level of those in the control group. As Pfannkuch et al. explain, there are two possible explanations for the observed difference.

1. The variability can be entirely explained by chance factors (differences between individuals who happened to be randomly assigned to each group and measurement errors) alone. That is, in this example, the fish oil diet is no better than the regular oil diet.
2. The variability required explanation by both chance factors and the treatment factor. That is, in this example, the fish oil diet treatment is effective (p. 905)

The researchers randomly reassigned the participants to one of the two groups and recorded the difference in the group means. After they repeated this process 1000 times, they noticed that the proportion of times the mean difference was greater than or equal to the observed difference of 7.71 mmHg was equal to 0.005. The authors concluded the observed difference is unlikely to be accounted for by chance alone and that the fish oil is in fact effective for reducing blood pressure.

While informal and simulation-based inferential methods appear promising for student understanding of hypothesis testing due to their simplicity and logical argumentation, it is important to examine empirical studies on their effectiveness. Below I summarize research examining the impact of informal inferential techniques on young children followed by studies investigating the effects of SBI on college.

Empirical Studies on Informal Inference

Young Children's Experiences with Informal Inference

Students can start learning informal methods of inference as early as in elementary or middle school. Several important themes emerged from the literature on young children's understanding of informal inference including (1) student understanding of variability, (2) the importance of selection and enactment of quality tasks, (3) the role

of context, (4) student reasoning about uncertainty, and (5) student use of language. These are described in detail below.

Understanding variability. Understanding of sampling distribution is fundamental for inferential reasoning. When students recognize how samples taken from the same population vary due to random chance, they are able to make better judgements about whether or not an observed result is unusual (Garfield, Zieffler, & Ben-Zvi, 2015; Wild, Pfannkuch, Regan, and Horton, 2011). Unfortunately, students experience difficulties when reasoning about variability between samples. Lee, Angotti, and Tarr (2010) examined how middle school students engaged in an informal hypothesis testing using a computer-simulated die-tossing experiment. They asked six students in an urban, public middle school to investigate the fairness of a die by simulating rolls of the die. Through observations and interviews, they found that the students generally reasoned about variability within samples (how frequencies of outcomes differed in one sample), rather than across samples (how distributions of outcomes varied from sample to sample). In rare instances, when students did notice how results across samples varied, they failed to keep the sample size consistent from one sample to another.

In a similar qualitative study involving two students aged 10-11 years, Pratt and colleagues (Pratt, Johnston-Wilder, Ainley, and Mason, 2008) noted that learners struggled with variability. As students simulated rolls of dice, they expected invariance in the data distribution, even with the small samples. As the researchers explained, "seeing repeated change seemed to makes some children distrust taking larger samples"

(p. 124). As a result, the students did not place a greater confidence on inferences based on large samples of data than those made from small samples. Another interesting result was that the students searched for stability at the wrong level. They expected that a different set of data would produce similar results (invariance across samples), rather than achieving stability with the increase of the sample size.

It should be noted that in both studies the participants were not given any directions about how much data to simulate or which graphical representation to use. Perhaps, entry-level statistics students should be guided on how to conduct data-driven investigations until they gain experience with such tasks. This recommendation is supported by Makar and Rubin's (2009) study, in which learners were asked to compare handspan data from their class to the data of their neighboring class. Based on the similarities and differences between the two data sets, the teacher encouraged and guided the learners to make predictions about the distribution of handspan data in a third class. This activity exposed students informally to the idea of variability between distributions and provided a foundation for drawing inferences.

Selection and enactment of quality tasks. The selection and implementation of a task is extremely important for the development of students' statistical reasoning. In their qualitative study of twenty-two 8-year-old students, Paparistodemou and Meletiou-Mavrotheris (2008) found that students were highly engaged with the task of collecting and analyzing data about health, nutrition, and safety habits of students in their school. They were eager to learn more about their school and often made conjectures based on

their personal experiences. For example, they found that older students tend to play less with scissors at school and explained this result by arguing that older students understand better that playing with scissors could be dangerous. Similarly, they found that boys in their school exercised more than girls and explained that this is because more boys participate in sports. Since the students were familiar with the context of the problem, they were able to understand and explain results.

Makar and Rubin (2009) also stressed the importance of choosing age-appropriate, relevant, and engaging tasks in their study mentioned above. One of the teachers in their study asked her fifth-grade students to investigate the opinions of children and adults about their views of Australian rights of citizenship. Her students struggled to make any conclusions based on their collected data. The teacher later reflected that students were not interested in the topic because it was not relevant to them. In addition, the data was dichotomous and lacked sufficient complexity to be used for making generalizations.

Not only selection but also implementation of a task affects the development of student statistical reasoning. In Makar and Rubin's (2009) study discussed above, the teacher did not pose a driving question. Therefore, the students did not have a clear purpose of making generalizations and predictions from their data. Instead, they focused on summarizing and describing data. On the other hand, when another teacher guided students to think beyond their collected data, they moved from focusing on individual data points to thinking holistically, considering the context and the problem of inquiry.

Holistic or global thinking is typically challenging for students. In the study

described earlier, Pratt et al. (2008) reported that the participants in his study were more concerned with individual trials than with the aggregated long-term outcomes. Perhaps students were not able to think globally because they were not guided through the investigation process and because their die-tossing task was not as relevant to them.

Context. Context is an important factor in developing students' inferential reasoning. Students often have difficulty focusing on the context under investigation. They treat data as an isolated artifact and therefore are unable to link their conclusions to the data and make generalizations (Makar and Rubin, 2009).

Pfannkuch (2011) distinguished two types of context that affect IIR: data-context and learning-experience-context. Data-context is the context of the real-world situation in which the problems arose, including the subject matter knowledge, how the data were obtained and how variables were measured. Learning-experience-context involves the prior statistical knowledge students bring to class. Based on her qualitative study of 29 tenth-grade female students, the author concluded that while a data-context can assist learners in understanding and explaining patterns in their data, it can also interfere with the reasoning process. For example, the data-context can divert student attention during the construction and application of concepts as learners begin to attend to context more than to properties that data reveal. The students in Pfannkuch's study became engaged with explaining possible reasons for having an outlier in their data and did not pay attention to variation between medians, which hindered their understanding of the sampling variability concept. On the other hand, the learning-experience-context played a

crucial role in developing students' IIR. The author concluded that "instruction may need to deliberately suppress the data-context at salient moments" (p.43).

Other researchers believe that students develop their statistical reasoning through an interaction between their contextual knowledge about the data set and their background statistical knowledge (Makar, Bakker, & Ben-Zvi, 2011; Neumann, Hood, & Neumann, 2013). Teachers need to help students coordinate contextual and statistical knowledge to help them make sense of a discrepancy between what they know from their prior experience and what they observe in data (Makar et al., 2011).

Reasoning about uncertainty. In the process of drawing conclusions about populations based on sample data, it is crucial to recognize the presence of uncertainty (Makar and Rubin, 2009; Manor, Ben-Zvi, & Aridor, 2014).

In their qualitative study, Manor et al. (2014) examined two 12-year-old male students' development of reasoning about uncertainty while engaging with informal inferential methods. The learners had to make inferences about a population distribution by examining random samples drawn from a population. The variables under interest were average time spent on social networks and number of friends in social networks.

The authors identified two types of uncertainty in the students' expression: contextual and statistical. The contextual uncertainty originated from a conflict between data and the students' contextual knowledge. When the boys examined a sample size of 10, one of the students found it unbelievable that a fourth-grade student had 600 friends in his social network, which made him resistant to make a conclusion from such a small

sample. On the other hand, statistical uncertainty stemmed from when the students observed variability in the data and variability between sample means. This phenomenon made the boys uncertain about inferring from a single sample of a specific size.

The authors concluded that "contextual uncertainty drove the boys to refine their methods of examining, controlling, and quantifying the statistical uncertainty" (p. 6) as the students transitioned between the following four stages: (1) Account for uncertainty, (2) Examine uncertainty, (3) Control uncertainty, and (4) Quantify uncertainty.

Research shows that young students are able to make generalizations and at the same time realize that their conclusions might be wrong. Paparistodemou and Melatiou-Mavrotheris (2008) studied twenty-two 8-year-old students in Cyprus to investigate how young learners begin to reason about informal inference by collecting and analyzing data about health, nutritional, and safety habits of students in their school. The study revealed that the students often went beyond their data and made general conclusions about all children in Cyprus while recognizing that their conclusions were not certain. Using probabilistic language, such as "more likely", "might be", "more possible to," etc., allows students to take risks in making their predictions without worrying about being wrong (Makar & Rubin, 2009).

Use of language. Language has an important role in developing students' statistical reasoning. Pfannkuch, et al. (2010) expressed that almost all pedagogical attention is focused on technical capability and not enough attention is paid to the integration of statistical concepts, extracting meaning, and increasing students'

communicative capability. The researchers urged educators to encourage students' use of language to "tell stories" from data. However, the language used might cause confusion. The authors claimed that the use of dense, technical jargon often creates misunderstandings in students, whereas natural, everyday language might lead to ambiguity. Nevertheless, the authors suggested to "operate as closely as possible to natural language but with some compromises around shared meanings" (p.3), such as "center," "shape," "spread" etc. The authors called these terms "loosely defined pieces of jargon," which are in between technical terms and natural language. We adapted these terms from everyday language "to convey the essence of a few very broad ideas" (p.3).

Caution should also be used with the use of metonymies, which occurs when students or teachers use approximate, but closely associated, terminology with a given concept. In their qualitative study, Null and Hancock (2015) showed that the use of metonymies can help but also sometimes hinder students' understanding of inference. While using their own terms might help students express their ideas quickly, without articulating every detail, it can often result in misconceptions. For example, when using the term "collection of many samples" for sampling distribution, students might incorrectly conclude that sampling distribution is formed by combining many small samples to form a large sample. The authors also warned that teachers should be cautious when they naturally shorten sentences without emphasizing the details of statistical terms. They need to give careful attention to concepts and their meanings, especially when they are first introduced.

In summary, statistics instruction can promote the development of students' inferential reasoning at an early age when students are (1) engaged in a variety of authentic activities related to their interests, (2) guided in the investigation process, (3) encouraged to use appropriate language and articulate uncertainty, and (4) prompted to pay attention to variability within and between samples.

Effectiveness of Simulation-Based Inferential Methods

The literature on simulation-based methods of inference is sparse and is summarized below.

Effectiveness of simulation-based curricula. As stated earlier, several researchers reformed and implemented a traditional introductory statistics course to include simulation-based inferential methods throughout the course. In their curriculum redesign, Tintle et al. (2011) deemphasized descriptive statistics and only covered probability and sampling distribution concepts implicitly while teaching randomization tests. They compiled the new curriculum into a textbook titled *An Active Approach to Statistical Inference* (Tintle, VanderStoep, &Swanson, 2009) and implemented it in eight sections of statistics course at Hope College, located in Holland, Michigan. The course was taught in a computer lab following an active-learning approach. The researchers administered the CAOS (delMas, Garfield, Ooms, & Chance, 2006) test to 202 students before and after the course. They compared results to the scores of 195 students who took the same course using a traditional curriculum two years earlier. In addition, they

compared the results to a nationally representative sample of 5,362 students who used a traditional curriculum and took pre and post versions of CAOS assessment.

The authors found that while the aggregate CAOS scores were similar across the three groups, there were significant differences in some topics. Students following the new curriculum outperformed the other two groups in questions regarding tests of significance, design, and simulation. However, they had poorer performance on estimating standard deviation from boxplots.

The study had several limitations. First, the test-taking conditions were different across the samples. Students in the new curriculum group took the tests at home with no performance (only completion) incentive, whereas the traditional curriculum students took the tests in a computer lab and were offered a performance incentive on the post-test. Second, the treatment consisted of the change in the curriculum as well as in pedagogy. Therefore, it is unclear which of these two contributed to the difference in scores. Finally, the CAOS assessment consists of multiple-choice questions with 2-4 response options. Therefore, many answers might be based on chance rather than on true understanding. Open-ended questions or student interviews would have provided more in-depth information on student thinking and would have strengthened the study.

Similarly, Garfield, delMas, and Zieffler (2012) tested their CATALST curriculum on 102 students enrolled in four sections of introductory statistics course at the University of Minnesota (78 students) and at North Caroline State University (24 students). Three assessment instruments were developed, two of which, (GOALS and

MOST) measured students' statistical thinking and reasoning and one assessed students' attitudes toward the course. It should be noted that reliability and validity of these instruments were not studied.

The researchers found that the majority of students felt the course helped develop their ability to think statistically and understand statistical information they hear or read in the news. The students performed well on questions involving graphical representations of data and the use of a randomization test, but struggled with reasoning about the effect of increasing sample size on test results. They had the most difficulty on the items related to testing a claim about a percentage.

The researchers compared the students' performance on GOALS items to the data from a national sample of 5,362 students who completed similar items on the CAOS test and found that CATALST students outperformed non-CATALST students on nine out of twelve items and did slightly worse on two items. The analysis was not rigorous and involved only descriptive statistics. In addition, the authors did not specify on which questions they found differences in student performance, making the results unclear.

Maurer and Lock (2016) criticized the above-mentioned studies for their weak design (the lack of random assignment) and non-rigorous analytical methods. They randomly split 101 students at Iowa State University into treatment and control groups. The treatment group covered simulation-based inferential methods (for both confidence intervals and hypothesis tests) in addition to formal inference, whereas the control group students were only exposed to formal approaches to inference. Based on the Assessment

Resource Tools for Improving Statistical Thinking (ARTIST, 2006), the researchers concluded that while students in the simulation group had significantly higher learning outcomes for the topics related to confidence intervals, the two groups did not differ on hypothesis testing topics. The authors attributed this result to the fact that the treatment group students had to learn both, formal and informal, methods of hypothesis testing in the same amount of time.

The above studies indicate that simulation-based inferential methods have a potential for increasing students' understanding of concepts related to hypothesis testing. However, because the researchers assessed the curriculum as a whole and did not supplement their quantitative findings with qualitative results, the results are of limited value and do not depict a complete picture about the impact of such methods on student knowledge.

Retention. Only one study was found on student retention of statistical concepts after exposure to randomization-based methods of inference. Tintle, Topliff, VanderStoep, Homes, and Swanson (2012) compared the conceptual understanding of a cohort of 79 students who used a traditional curriculum to a cohort of 76 students who took a randomization-based curriculum. Students in each group completed the CAOS assessment at three separate times: during the first week of class, during the last week of class, and four months after finishing the course. After controlling for pre-test and post-test scores, a multiple regression model revealed a significant difference ($p = 0.002$) in aggregate retention between the two groups, with the randomization group having higher

retention. The largest difference was found in the areas of data collection/design and statistical tests (logic and interpretation of *p*-value). The authors attribute these differences not only to curricular changes, but also to active-learning pedagogy that incorporated more projects, lab exercises, group problem solving, and discussion compared to the traditional group.

It should be noted that only 40% of the students took the test four months later and the students self-selected, weakening the validity of the results.

SBI terminology. When teaching simulation-based methods, it is important to pay close attention to the language used in describing concepts. Pfannkuch et al. (2011) demonstrated the randomization method for comparing two means. They re-randomized participants into the treatment and control groups 1000 times and constructed a dot plot depicting the mean differences between groups. The authors argued that the graph cannot be called a "sample distribution" because the data were not obtained by drawing samples from the population. In addition, they called the process of random re-assignment of units into groups "re-randomization," which could be misleading for observational studies, in which the participants are not randomly assigned to groups from the beginning. They also suggest using the term "tail proportion" for describing the proportion of re-randomizations that produced a difference in the group means at least as big as the observed difference. Although the concept of tail proportion is similar to the notion of the *p*-value, the authors suggest not using the latter term because not all possible random allocations were generated when producing the dot plot. Finally, the authors claim that

randomization methods are often referred to as re-sampling methods, which is misleading because sampling is not occurring in the data production process.

Student difficulty with language and misuse of terminology was documented in several of studies. For instance, Budgett et al. (2013) observed that students struggled to understand double negatives in statistical conclusions. The authors proposed rephrasing the statement "I have no evidence that she has not brushed her teeth" with "I am still not sure if she has brushed her teeth" (p. 15). Moreover, students often confused the terms random assignment and random selection. However, their oral responses indicated that they understood the purpose of random assignment and simply misinterpreted the language of multiple-choice questions. Likewise, Pfannkuch and Budgett (2014) reported that a student in their study incorrectly used the term "resample," instead of "re-randomize" for describing random reassignments of participants in treatment and control groups. In spite of difficulty with terminology, the student was able to correctly explain each stage of the randomization method and interpret the results.

In summary, language is an important part of understanding and expressing statistical ideas. Student responses on multiple-choice assessments might not reveal their true understanding of concepts and, if possible, should be accompanied with interviews and oral clarifications.

Hands-on vs. computer-based activities. Many researchers suggest engaging students in tactile activities, such as simulating binomial distribution with a coin flip or randomizing subjects into groups by shuffling cards, before exposing them to computer

simulations. Fitch and Regan (2014) claim that "the randomization method lends itself to tactile and visual experiences" (p.3). Budgett et al. (2013) also believe that "the hands-on component is a critical part of the conceptual understanding of the randomization test and that without it students will have difficulty in reconstructing the process when faced with a new scenario" (p. 16). Many of their students indicated that the hands-on activity helped them understand the components of the computer program. The researchers suggested using the combination of technology and hands-on activities in order to facilitate student understanding of the randomization method. However, Holcomb and colleagues (Holcomb, Chance, Rossman, Tietjen, and Cobb, 2010) provided contradictory findings.

They divided 43 students into two, "tactile followed by simulation" and "only simulation" groups and exposed them to a task involving simulation of sampling distribution about proportion. The students then were given a quiz that tested their understanding of concepts related to the task. The researchers did not find significant differences between the two groups on student performance on the quiz. However, about half of the students in the tactile group commented that the hands-on simulation with card was beneficial because it "helped them understand what the computer was doing, involved them in the process, and … is better for visual learners" (p.6).

Students' interaction with dynamic visualizations. Students' experiences with computer simulations for statistical inference are mixed. When Rossman and Chance (2014) piloted their revised curriculum incorporating simulation- and randomization-

based inference, they observed that some students struggled to understand the purpose of the simulation and incorrectly believed that it was used for replicating a research study. Other students did not understand what a simulation was and thought that it referred to any piece of technology. Learners also had difficulty understanding what observation units and variables were in a randomized distribution and struggled to identify the parameter of interest. Likewise, Gould, Davis, Patel, and Esfandiari (2010) reported student problems with basic computer operations, such as being able to find downloaded files, or perform data cleaning procedures, which created frustration in students and distracted them from understanding simulations.

On the positive side, many researchers uncovered benefits in using computer simulations for understanding inferential concepts. Budgett et al. (2013) reported all ten participants in their mixed-methods study understood every component of the program and were able to explain what each one represented. The authors advised using software in combination with verbalizations and language to help students understand and reason about simulated data. Similarly, Lipson, Kokonis, and Francis (2003) found their simulation effective. The researchers were guiding students during their interaction with the software. They discovered that initially students in their study had difficulty interpreting the sampling distribution. However, after going back to generating the empirical sampling distribution, the learners understood how the sampling distribution was obtained, which the researchers attributed to the dynamic nature of the simulation. The authors recommended accompanying simulation activities with a series of questions

in order to ensure that students understand and pay attention to all of the components of the screen display.

In conclusion, a computer-program alone is not likely to promote student understanding of statistical concepts, unless it is scaffolded with directions and explanations.

Student understanding of inference with simulations. Studies on student reasoning about concepts of hypothesis testing using the randomization and resampling methods are limited. Lipson et al. (2003) studied how a group of eight students interacted with a computer simulation activity to explore the concepts of hypothesis testing using binomial simulation prior to learning about formal methods of inference. The students were given a real-life example, in which the Australian postal service claimed that at least 96% of its letters were delivered on time, while sample data collected by a journalist suggested that the on-time delivery rate was in fact equal to 88%.

Student interviews revealed that in order for them to understand the logic and concepts of inference, the learners had to move through the following four stages: recognition, reconciliation, contradiction, and explanation. During the first stage students interpreted and connected various screen representations and constructed their schema for empirical sampling distribution for the proportion. In the reconciliation stage students linked the journalist's sample results with the sampling distribution, and realized that the journalist's sample was unlikely. In the third stage, students observed the contradiction between the journalist's result and the post-office's claim. Finally, the fourth stage

involved consideration of the possible statistical explanations for the contradiction between the observed sample and the hypothesized sampling distribution.

Students struggled the most in the final stage. They used absolute terms, such as "the Australia Post lied," rather than statistical explanations to interpret the conflict between the journalist's sample and the hypothesized population. On the other hand, some students did not reject the post-office's claim in spite of the small probability of the journalist's result. The authors explained that the context of the problem played a role in distrusting the sample results. The students viewed Australia Post as a large and reputable organization, while the journalist was perceived as a corrupt individual.

In their mixed-methods study, Budgett et al. (2013) also explored student understanding of statistical concepts with the randomization method using dynamic visualizations. Ten high school and university students of varied backgrounds were given a two-hour teaching session on randomization tests and then were interviewed and tested one week later. The researchers found that many students considered 10% as a strict cut-off between significant and non-significant results. In fact, one student thought that the tail proportion was a measure of the effectiveness of the treatment, tail proportion of about 50% indicates that the treatment is not useful, around 27% somewhat useful, and less than 10% useful.

On a positive note, all of the students were able to explain how re-randomization distribution was obtained, which was a problem with Gould et al.'s (2010) participants, who thought that re-randomized distribution was the observed data distribution.

Another controversy deals with student understanding of significant and non-significant results. While some researchers showed (Pfannkuch & Budget, 2014) that students are equally capable of interpreting both significant and non-significant findings, others (Holcomb et al., 2010) demonstrated that "students find it easier to spot a surprising outcome than a non-surprising one" (p.5). The latter statement was confirmed by Budgett et al. (2013), who found that most of their students incorrectly interpreted a large *p*-value as "evidence in favor of the chance-alone explanation" (p. 13). The learners applied an indirect logic, reasoning that if small tail proportions provide evidence against the null model, then large tail proportions will give evidence in favor of the null model. However, this logic is false because while a large tail proportion provides no evidence against chance acting alone, there might be some other factors acting along with the chance.

In summary, while some students are able to grasp simulation-based inferential methods, others experience misconceptions, such as, significance level is a strict cutoff score between significant and non-significant results, a large *p*-value indicates that chance is acting along, re-randomized distribution is the same as the distribution of the observed data, and conclusions based on inference are deterministic. Clearly, more research is needed to understand student reasoning with informal, simulation-based methods and what is the source of student misconceptions.

The Present Study

As shown above, the research on student understanding of randomization and permutation tests is scarce and conflicting. While there are several studies examining student understanding of simulation-based inferential methods (Budget et al., 2013; Garfield et al., 2012; Holcomb et al., 2010; Lipson et al., 2003; Pfannkuch & Budget, 2014; Tintle et al., 2011), they are not in-depth and lack rigor. The studies either rely on multiple-choice questions (Holcomb et al., 2010), report results for only one student (Pfannkuch and Budgett, 2014) or one activity (Lipson et al., 2003), or assess the impact of curricular changes as a whole, lacking qualitative support (Garfield et al., 2012; Tintle et al., 2011). In addition, none of the presented studies explore teachers' interpretation of and views on simulation-based inferential methods. Finally, no studies were found examining community college students' experiences with SBI.

This project will attempt to close the above-mentioned gaps in the literature and examine in detail students' and teachers' reactions to SBI in a community college setting.

Research Questions

The study will investigate the following research questions:

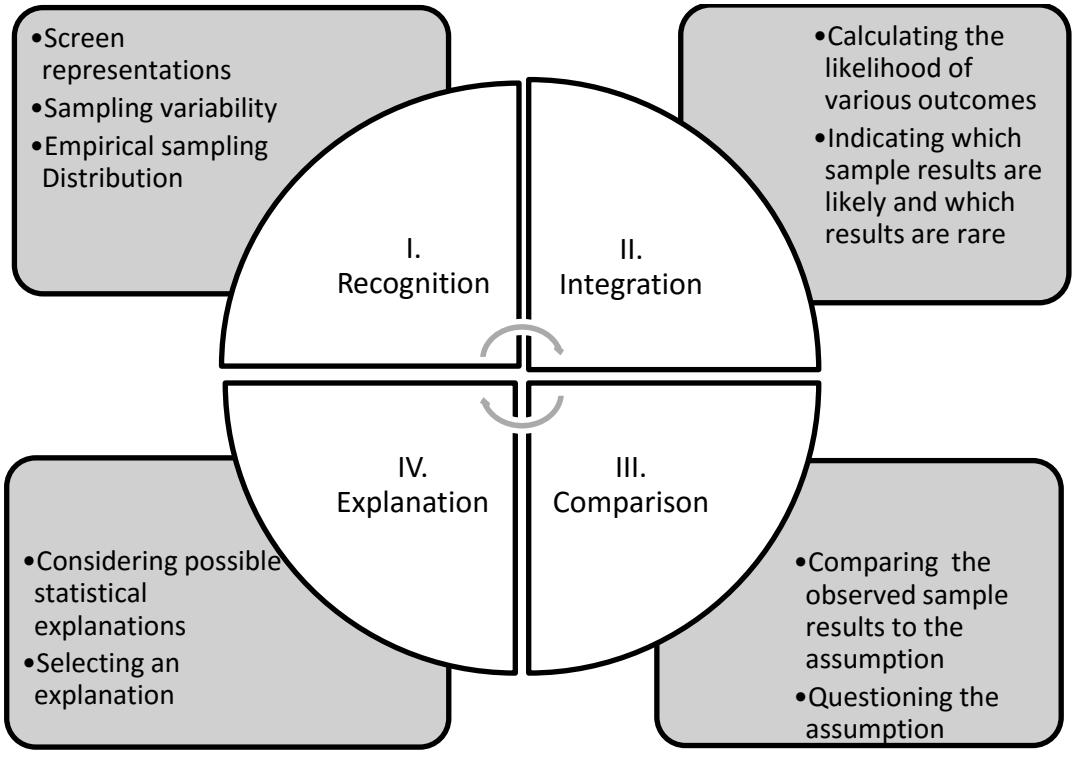
1) How do instructors describe their students' understandings and misunderstandings of inference?
2) What are instructors' attitudes toward simulation-based inference?
    a. How do instructors compare formal and informal inferential methods as instructional approaches in introductory statistics courses?
    b. How do instructors compare formal and informal inferential methods as statistical methods?
3) What do community college students who took an introductory statistics course recall and understand about formal hypothesis testing?
4) How do community college students with and without a statistics background understand the functionality and logic of the simulations?
    a. What do students understand and what difficulties do they experience when learning inference through a simulation-based approach?
    b. To what extent do students who had exposure to formal inferential methods make connections between informal and formal inferential methods?
    c. How do students who had exposure to formal inferential methods compare and contrast formal and simulation-based inferential approaches?

Conceptual Framework (adopted from Lipson et al. (2003))

The conceptual framework for this project is based on Lipson et al.'s (2003) qualitative study that examined students' interaction with a computer simulation designed to support their understating of hypothesis testing. The authors found that students moved through the following four stages in the process of understanding the logic of inference: Recognition, Reconciliation, Contradiction, and Explanation. I slightly modified the second and third stages to fit the purpose of this project and depicted them in the visual below (see Figure 2).

Figure 2

Conceptual Framework



According to Lipson et al. (2003), the recognition stage involves three substages: (1) interpretation of the components of the computer screen, (2) understanding that sample results vary even though they are taken from the same population, and (3) interpretation of the empirical sampling distribution.

The authors suggest that at the second stage students link the observed sample results with the hypothesized sampling distribution and calculate the likelihood of the occurrence of the observed sample results. However, I would like to observe first whether students are able to compute the likelihood of various possible outcomes of the empirical distribution and to distinguish common results from those that are improbable.

I renamed the third stage from “contradiction” to “comparison.” Students at this stage compare the obtained sample results to the initial assumption. Since I want to observe how learners explain both extreme and likely outcomes, there is not always a contradiction. Instead, students should decide whether they have enough evidence against the null hypothesis based on the likelihood the obtained results.

At the final stage students consider the possible statistical explanations for their observed results and formulate their final conclusion.

3. METHODOLOGY

Research Design

Rationale for Qualitative Methods

This study investigated how students and instructors reason about simulation-based inferential methods. Since the questions focus on process rather than on variance, qualitative methods are most appropriate (Maxwell, 2005). Marshall and Rossman (2010) emphasize the strengths of qualitative methodology for eliciting tacit knowledge and interpretations of participants and for in-depth exploration of the complexities of a phenomenon. The authors argue that, since human actions are significantly influenced by the settings in which they occur, the studies should take place in real-life situations. In addition, qualitative methods allow researchers to collect data in natural settings and uncover multiple versions of reality.

According to Maxwell (2005), "The strengths of qualitative research derived primarily from its inductive approach, its focus on specific situations or people, and its emphasis on words rather than numbers" (p.22). The present study focused not only on whether SBI improves student statistical reasoning, but on how students interpret and justify each step of the hypothesis testing process. Moreover, I examined how participants make meaning of statistical inference in the context of computer simulations, which directly fits a qualitative research approach. Specifically, I employed a case study design. According to Yin (2014), case studies are appropriate when the research questions depend on the contextual settings of the events. Since the questions and

findings of this study depend on the school settings, case study design was most favorable.

Research Paradigm

I approached this study through an interpretive paradigm lens. I believe that people construct their realities and make their own meanings based on their experiences. This paradigm assumes that the researcher and participants are interactively creating findings during the study. According to Guba and Lincoln (1994), "constructivism's relativism assumes multiple, apprehendable, and sometimes conflicting social realities that are the products of human intellects, but that may change as their constructors become more informed and sophisticated" (p. 111). In my view, teachers and students make their own meaning and interpretation of the SBI. In addition, I as a researcher was part of the meaning-creating process.

Site and Methods

Site

I collected data from statistics instructors and students at a mid-sized community college in Central Virginia (CCCC). According to an institutional research report published by the college in 2016, the student population consists of 79% part-time and 21% full-time students. 59% of the students are female while 41% are male. The average age of the students is 22.8 years.

Participants

The participants consisted of four statistics instructors and six students. All instructors are full-time faculty at CCCC with a rank of Instructor of Mathematics. They were chosen for the study because they have previously taught elementary statistics. Their education and teaching experience is outlined in Table 4.

Table 4

Instructor Information

	Scott	**Morgan**	**Cassidy**	**Clark**
CCCC Teaching Experience	2.5 years	2 years	5 years	10 years
Total Teaching Experience	14 years	23 years	21 years	35 years
Statistics Teaching Experience	2.5 years	2 years	2 years	35 years
Education	B.S. in Mathematics, M.Ed	B.S. & M.S. in Mathematics	B.S. in Mathematics, M.Ed	B.S. in Mathematics, M.S. in Statistics

Six students participated in the study, four of which completed an introductory statistics course at CCCC while two students had no previous statistical background (see Table 5).

About 250 CCCC students were contacted via email and asked to volunteer in this study (see email in Appendix B). Approximately half of these students have previously taken an introductory statistic course at the college. Sixteen students expressed interest in

participating, seven of which had completed statistics within the last year. I chose four students with a statistics background and ensured that they were diverse in terms of grades in the statistics course. Out of twelve students who had not taken statistics at CCCC, ten were excluded due to one of the following reasons: (1) they were younger than 18 years, (2) they took statistics in high school, (3) they either did not respond to emails with scheduling requests or did not show up for the interview.

Table 5

Student Information

	Lena	**Natalie**	**Casey**	**Derek**	**Ethan**	**Anne**
Age	21	19	20	21	20	23
Statistics Grade	A	B	C	D	-	-
Degree Program	Business Admin.	General Studies	Nursing	Education	Music	Nursing

Research Methods

The research methods consisted of teacher and student interviews, each having the following three phases: 1) pre-simulation interview, 2) exploration of SBI, and 3) post-simulation interview. Student and instructor interviews were audio-recorded and transcribed within 24 hours from the interview. During student interaction with simulations in addition to voice recording, the computer screen was captured via a video platform, *Panopto*.

The main method of data-collection consisted of in-depth interviews. Marshall and Rossman (2010) argue that in-depth interviewing is an effective way for gauging participants' views on the phenomenon of interest as it helps uncover their emic

perspectives. Interviews provide rich data because they allow for immediate follow-up and clarification. According to Patton (2002), interviews are most appropriate when we need to learn about participants' feelings, thoughts, and intentions, which are difficult to uncover from other data-collection approaches. Specifically, I utilized task-based interviews, which allowed me to ask probing questions in order to gauge the participants' understanding of the logic and functionality of the simulations through their explanations.

The interview questions can be found in Appendix A. Below I describe data-collection methods in detail.

Instructor interviews. Each instructor interview lasted about an hour and consisted of three phases described below.

Pre-simulation interview. During the first phase the instructors were asked about their experiences with teaching statistics and their views about their students' understanding of inference. The purpose of this phase was to gauge the instructors' attitudes toward teaching statistic and to uncover their perceptions of students' understandings and challenges related to statistical concepts.

Exploration of SBI. At the second phase the instructors were shown examples on how to use simulations to make inference about population proportion and to compare two population means. They were also asked probing questions to gather their understanding of SBI.

Post-simulation interview. After exposing instructors to informal statistical inference through simulations, they were asked to share their reactions toward

simulations, their views on informal statistical inference, and whether teaching SBI is likely to be beneficial for students. In addition, the instructors were asked for their recommendations about implementing SBI in introductory statistics courses.

Student Interview. Each student interview lasted 65-75 minutes and consisted of three phases described below.

Pre-simulation interview. Those students who took a statistics course were asked about their statistics experience and what they remembered about hypothesis testing. The questions were open-ended and inquired about their understanding of the purpose of hypothesis testing in general and of specific concepts related to this topic. The purpose of this phase was to gauge students' current knowledge of hypothesis testing before exposing them to simulations.

Guided exploration. After assessing students' understanding of statistical inference, I introduced them to binomial simulations and randomization tests. Those students who have not taken statistics were directly presented with simulations. Research suggests scaffolding students through the investigation of computer-based simulations methods to ensure that they can understand and verbalize every step of the process (Budgett et al., 2014; Lipson et al., 2003). I walked students through the stages described in the conceptual framework, assessing their reasoning at each stage.

Students were presented with two examples to help them explore informal inference. In the first example, students were guided to investigate whether a coin is fair if 10 tosses produced nine heads. In the second example, they explored whether fish oil is

effective in reducing blood pressure using data from Knapp and Fitzgerald's (1989) study described earlier.

The learners were also presented with a task that required them to transfer their understanding of binomial simulation from coin flipping to a more complex, but also more realistic, context. They were asked to use the binomial simulation to decide whether drug Y, which produced an improvement of 12 out of 20 patients, is more effective than drug X, which has an improvement rate of 40%. The purpose of this task is to assess whether students understood the binomial simulation and can apply it to a different task.

Post-simulation interview. After student exposure to simulations, I inquired about their views on SBI. In addition, I asked students who took a statistics course whether they prefer learning inference with simulation-based approach, with formal methods, or with a combination of the two approaches.

Computer Simulations

I used simulations developed by Rossman and Chance (2009A, 2009B). The researchers argue that technology should be easy to use so that students can focus on statistical ideas rather than on technological issues (Rossman & Chance, 2014). Below I describe binomial simulation and randomization test for two quantitative variables.

Binomial simulation. The binomial simulation applet (Rossman and Chance, 2009A) allows to make inference for a proportion. Students enter the null value to be tested, the sample size, and the number of repetitions to be simulated. They can choose which statistic to analyze: sample proportion or number of successes. They can calculate

an approximate *p*-value from the simulations results by entering the observed number of successes in the relevant box. The binomial simulation directly resembles by-hand simulation because the applet displays coins being tossed when the null probability is one-half, and displays spinners for the other values of the null hypothesis. Figure 3 depicts 1000 simulations of 10 tosses of a fair coin. It also shows that only five out of 1000 experiments result in 9 or 10 heads.

Randomization test. The randomization test for quantitative response applet (Rossman and Chance, 2009B) allows students to compare two groups with numerical responses (see Figure 4). Users can input their own data and calculate descriptive statistics for each sample. They can choose to compute one of the following statistics: difference in means, difference in medians, mean absolute difference, or range. The program allows to re-shuffle the original responses, compute the difference between desired statistics and construct the distribution of the differences. Finally, students can calculate the proportion of the observed or more extreme difference and draw a conclusion about the likelihood of the observed results.

Pilot Work

I conducted two pilot interviews with students in August, 2016. The purpose for these interviews was to practice my interviewing skills and to refine the tasks as well as the interview questions. The chair of my capstone committee attended the interviews and provided suggestions on the wording and sequencing of the questions, which I took into consideration for the actual interviews. For instance, during the first pilot interview I

presented the student with complete tasks from the beginning, which created confusion. We found that it was more beneficial to break the problems down. For example, it was preferable to first ask students to generate a sampling distribution for drug X and then ask the question about the effectiveness of drug Y.

During the pilot interviews, we also noticed that students were confused with the wording “reduction in blood pressures” presented in the randomization task. It took students time to understanding that as a positive reduction means that blood pressure dropped. As a result, we altered the wording “reduction” to “change” and switched the signs of presented numbers from positive to negative and vice versa.

Another benefit of the pilot interviews was to revise our conceptual framework described earlier. Finally, the pilot interviews helped me maintain the focus on my research questions instead of spending time on less relevant topics.

Figure 3

Binomial Simulation of a Fair Die

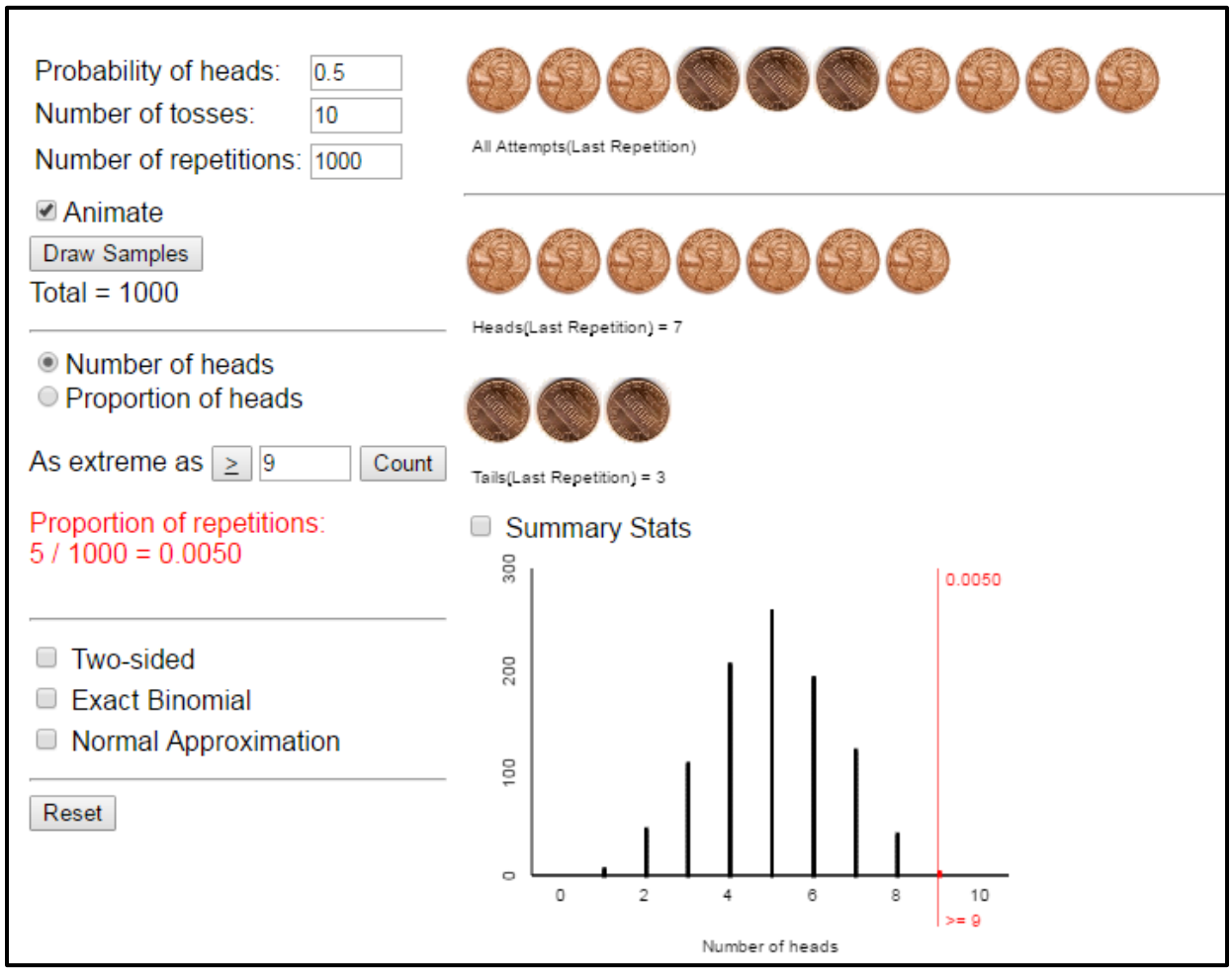
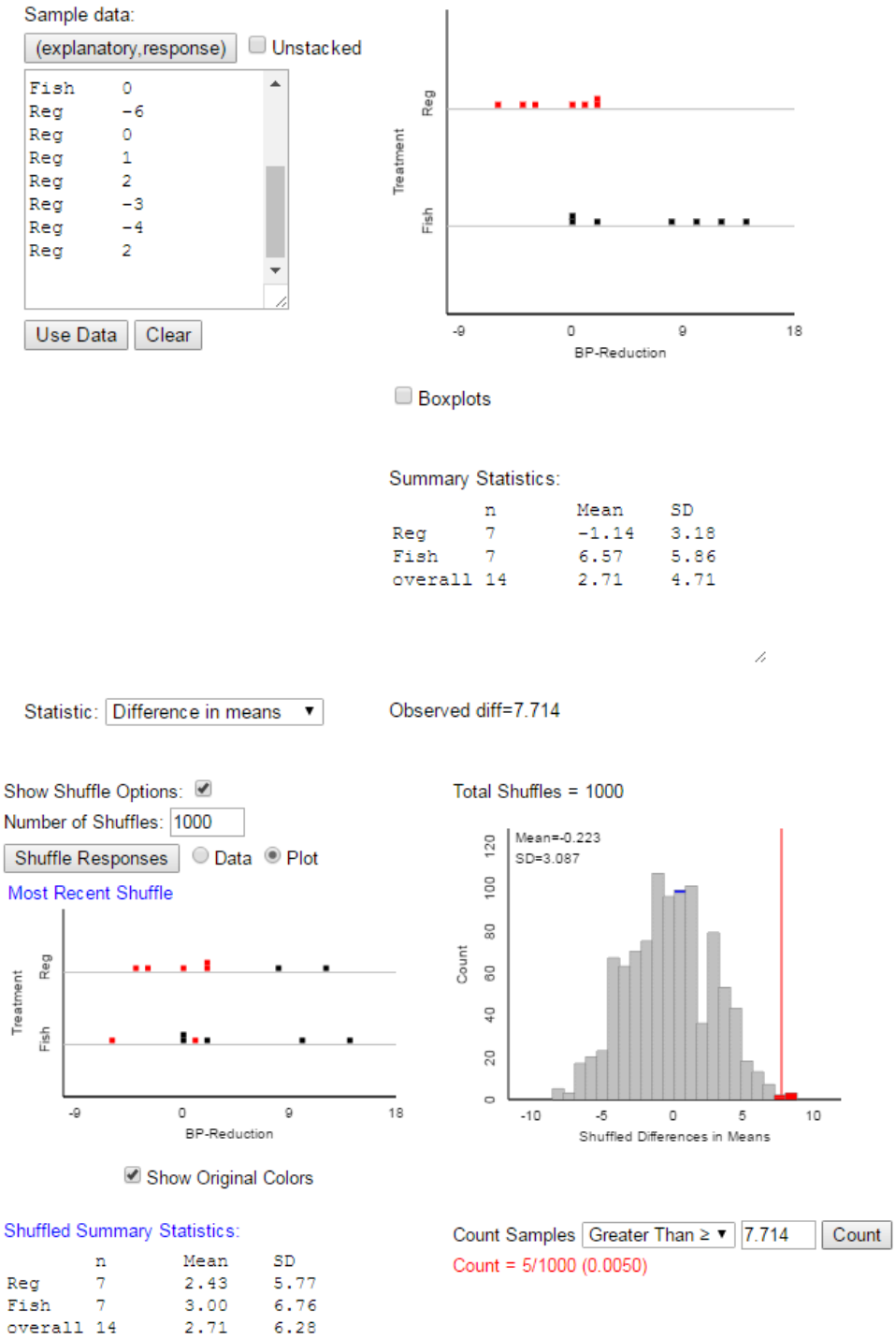


Figure 4

Randomization Test for Quantitative Response



Data Analysis

Data collection and data analysis occurred concurrently as advised by Miles, Huberman, and Saldana (2014). I employed Erickson's (1986) method of analytic induction that suggests creating assertions based on the data and repeatedly testing and modifying the assertions until they fit all instances. I proposed assertions through the process of memoing and reading interview transcripts multiple times. I then looked for disconfirming and confirming evidence and modified assertions when necessary.

I organized data in different ways – by cases, by research questions, and by interview questions. I created a matrix display of the data, which is useful for "condensing material into an 'at-glance' format for reflection, verification, conclusion drawing, and other analytic acts" (Miles et al., 2014, p. 91). The matrix display helped me observe patterns and themes in my data that I either verified or disproved by another round of data examination. Stake (2006) also proposed using a matrix to compare findings across cases and develop multi-case assertions. Student exploration of SBI was examined in a cross-case matrix over the four stages described in the conceptual framework: Recognition, Integration, Comparison, and Explanation.

I documented my reflections and thinking process about the data in my analytic memos. According to Miles et al. (2014), analytic memos are useful for synthesizing data into high level analytic meanings. They "tie together different pieces of data into a recognizable cluster, often to show that those data are instances of a general concept" (p. 96).

Validity

Researchers provide different checklists in order to deal with various validity threats (Guba and Lincoln, 1989; Maxwell 2005, Patton, 2000)**.** Maxwell (2005) warns that not every strategy works with a given study and urges the researcher to apply those strategies that are most feasible and efficient. In this study, I established validity by collecting "rich," detailed data that will be varied enough to provide a full picture of student and instructor reactions to and understanding of SBI. In addition, I "member-checked" the instructors' responses in order to validate their views. I followed up with each instructor to make sure I correctly interpreted their views on SBI and their recommendations on using them. Finally, I searched for disconfirming evidence and reported conflicting cases.

Ethical Considerations

I aimed to conduct this study in an ethically appropriate manner. I obtained access by applying to the Institutional Review Board at the University of Virginia. After the approval, I gained access to the site from the Director of Research at CCCC.

The participants were provided with a detailed explanation of the study and were given an informed consent letter (see Appendix B). I kept the identity of the participants and the research site confidential, and I provided anonymity through the use of pseudonyms. Participation in the study was voluntary, and students as well as instructors could drop out of the study at any point. Finally, all data were stored electronically on a password-protected computer.

Although I am in close collegial relationships with the instructors, I do not have any position of power over them. Moreover, I did not select any of my former or current students for participation in the study.

One possible risk is that the reader of study will identify the school and the teachers. In order to reduce this risk, I included a minimal information about the teachers.

Reflexive Statement

I am a full-time associate professor at CCCC where I teach statistics and mathematics courses. I have a master's degree is applied mathematics with concentration in statistics. Although I took five graduate statistics courses as part of my master's program, I gained a true understanding of inferential concepts only after I began teaching introductory level statistics courses. As an instructor, I have witnessed even high-ability students struggling with the concepts of statistical inference. While most of them are able to follow steps and perform computations, few of them truly understood the purpose and meaning of the concepts. Reading research articles about informal inference sparked my interest in trying this method of teaching. However, before radically changing my teaching approach, I decided to investigate the usefulness of simulation-based inferential methods.

I acknowledge that my experience, relationships with the participants, and my personal traits had an unavoidable effect on the data collection and interpretation. I tried to address my personal biases and preconceptions through reflexive journaling as suggested by Lincoln & Guba (1985).

4. INSTRUCTORS' EXPLORATION OF SBI

This chapter presents the analysis and results of instructor interviews. For the sake of anonymity, the names and pronouns used to describe the instructors do not reflect their genders.

Instructor Scott

Instructor Background

Instructor Scott has been teaching mathematics and statistics at CCCC for 2.5 years and has a total of 14 years of teaching experience. Prior to working at the college, Scott taught mathematics at a local high-school. He has taught a variety of courses ranging from arithmetic to differential equations and started teaching statistics about two years ago. Scott has an undergraduate degree in mathematics and a master's degree in education.

Attitudes toward Statistics

Professor Scott took an introductory statistics course in college but did not like it because "it seemed like it was just a rote memorization of formulas instead of the application of concepts." However, once he started teaching the course and gained a better understanding of concepts, Scott started enjoying it.

Scott reported three reasons for liking teaching statistics. First, he had not been exposed to this side of mathematics for many years and finds discovering new ideas enjoyable. Second, the instructor appreciates connections between statistics and calculus, which he was unable to make as a student. Scott frequently shows these connections to

students. For example, he demonstrates in class how to derive the slope of a regression line, which involves the use of derivatives. While most of the students have not taken calculus, Scott believes that it is important for students who took or will take calculus to be able to see how it relates to statistics. Finally, the professor appreciates that statistics is very applicable compared to an algebra course, in which "it's really tough to argue with a student that it's gonna be used in real life."

Scott especially likes teaching hypothesis testing because it is applicable and "lends itself to fun class examples." He spoke with enthusiasm about the activities they do in class before introducing this topic. Scott provides students with a deck of cards, in which 75% of cards are black. Students select cards one by one and soon start questioning the assumption of having a typical deck with 50% black cards.

Scott also spoke with fascination about the Central Limit Theorem saying, "It's amazing... It's beautiful… I love the fact that it does not really matter what you have to start with. It's gonna become a normal distribution." He expressed that at first the theorem does not appear intuitive, but after thinking about it for a while, it makes perfect sense. The instructor uses a computer simulation in class to help students visualize why the theorem works.

Scott does not enjoy teaching how to construct graphs and charts because "it's too simple and most students already know this." The only exception to this is introducing students to misleading graphs, which is interesting and new to them.

In summary, Scott appreciates the mathematics behind statistical concepts as well as their applications, and tries to explain to students the meaning of statistical ideas rather than simply teaching facts and procedures.

Perception of Student Understanding of and Difficulties with Statistical Concepts

Scott believes that students find the mathematics part of the course easy because an introductory statistics course does not require strong algebra skills. Students typically are able to perform computations and find the right answer. However, in Scott's view, students too often have difficulties interpreting their answers, and, generally, they struggle with language. He feels that inference is most problematic for students because the wording of the questions. For example, students have a hard time deciding which hypothesis test to use, whether they are comparing proportions or means, whether the samples are independent or dependent, etc.

Binomial Simulation

Exploring the simulation. As I was demonstrating the binomial simulation with the coin example, Scott quickly and correctly answered my questions, demonstrating an understanding of the program. He predicted an approximately normal empirical sampling distribution for a fair coin and the fact that obtaining nine out of ten heads when tossing a fair coin would be quite unlikely. Once we calculated the tail proportion, Scott remarked, "Look at the *p-value,* it is as low as we expected,*"* making connections between the simulation and the formal hypothesis testing.

When I presented the problem about drug X, Scott was able to set up the simulation correctly. He laughed when the program displayed spinners instead of coins and thought that it was a good visual. After computing the tail proportion of 0.046, Scott remarked, “I guess based on 5% cut-off, drug Y is better.”

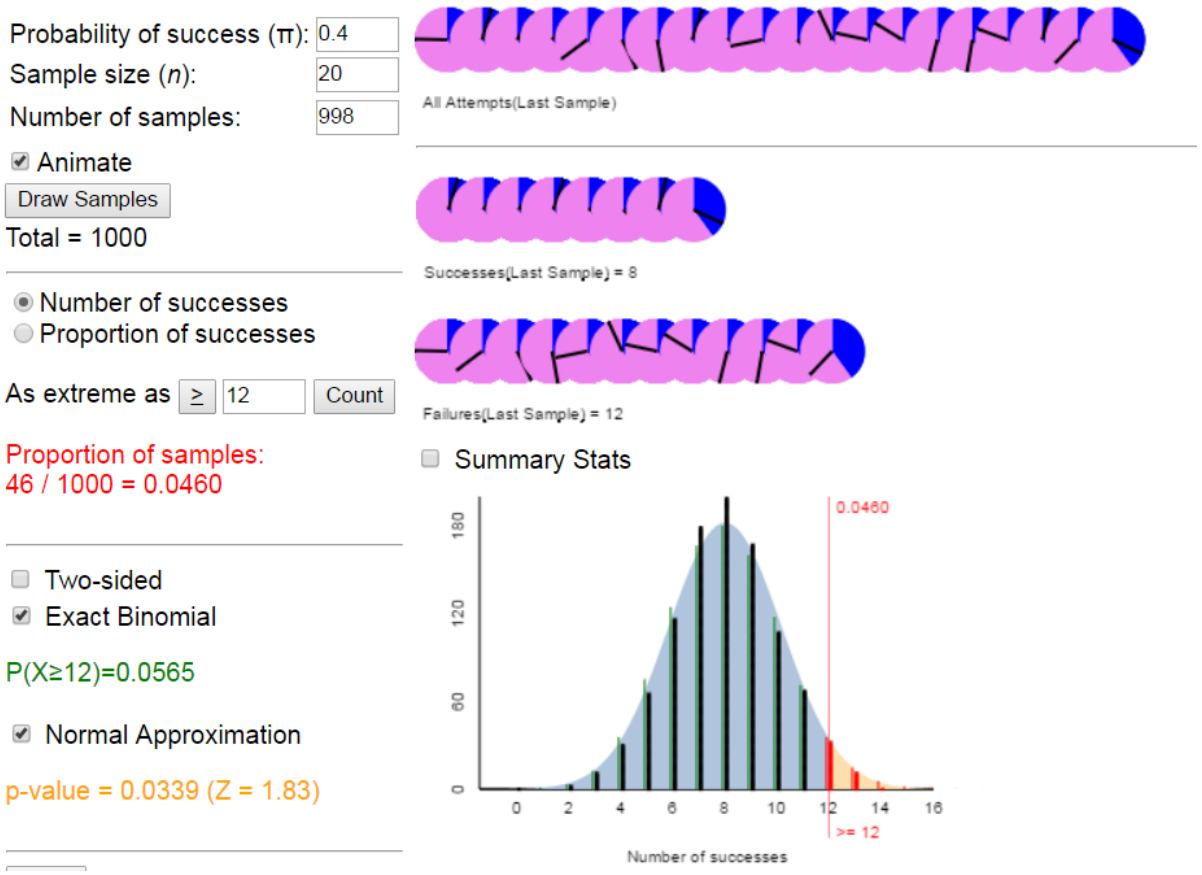
Scott was interested in comparing the simulation results, the exact binomial proportion, and the *p-value*. He was very surprised when we observed a large difference between the binomial probability of tossing 9 or 10 heads (0.0565) and the *p-value* using a normal approximation (0.0339) as shown in Figure 5. Scott exclaimed, “That’s so crazy! One is clearly reject and another is fail to reject!” Scott expected the two proportions to be closer to each other, saying “I explain to my students that a normal distribution is a decent approximation of the binomial distribution, but I haven’t seen this comparison… I did not realize they are so far!”

Reactions to the simulation. Even before I inquired about Scott’s views on the binomial simulation, he remarked that the program was helpful for explaining the meaning of *p-value* because often students confuse a *p-value* with confidence level. Scott elaborated,

> This is where students often get hung up. Because a lot of times, and I am sure you had that too, students get confused that a low *p-value* is good… they are thinking back like 98% confidence level, and they want *p-value* to be close to one…this [program] shows them what the *p-value* is. To me, seeing this and saying [to students], “It looks like if you run this experiment a bunch of times, you know you’re gonna get different results. This is how often you’re gonna get the result described in this problem. That seems kind of unlikely to me, how about to you?” I think it seems very tangible. So, I think it makes the concept easier to understand, especially probably the first time someone sees it.

Figure 5

Comparison of Tail Proportions



Scott also liked the binomial simulation because it clearly depicts that the sampling distribution is discrete. Scott noted that with a normal approximation, students often incorrectly assume that the sampling distribution is continuous. He remarked, "It's funny… They [students] have a hard time understanding discrete and continuous data,

and yet we only let them deal with continuous data for their cdf's [cumulative distribution functions]. That does not make sense."

Using the simulation for teaching. Scott expressed that the simulation would be especially useful for introducing students to hypothesis testing. When he first started teaching statistics one of the students had a difficult time understanding the *p-value* concept. Scott shared,

> I ended up calling my brother, who I knew had not had stats, but I also knew he was a smart guy, so I kinda wanted to explain this idea to him... Kind of run it by and see if it seems like it was tough for him or not. He ended up getting it, but it was hard [to explain]. With this I think he would have gotten it right away... It [the simulation] definitely makes it very visual. I love it!

In Scott's view, the binomial simulation is a great tool for introducing students to hypothesis testing. The instructor reported, "I would use this [simulation] to introduce the topic, and maybe come back to it every once in a while to reinforce the concept… and I will use it to compare the answers."

In summary, Scott viewed the simulation as a very helpful tool for explaining the concepts of hypothesis testing to students. He reported, "There is no way that I would teach this [hypothesis testing] again without showing this [binomial simulation]."

Despite Scott's positive views of the simulation, he did not feel comfortable completely abandoning the formal method for the following reasons. First, the professor was worried about the departmental final exam, which requires the use of formal methods. However, even when I asked him to put aside departmental restrictions and

imagine having the freedom to design a statistical curriculum, Scott still reported favoring a blended approach.

> I don't know… I don't love seeing how off the binomial and normal approximations are. Hm… I like the concepts behind what I am used to doing. But it makes you really wonder about standard error that we are using. And how off it is when we do it… I think that I would still do the two together.

It appears that Scott advocates for the blended approach because he appreciates the mathematics behind the formal method even though it does not always produce accurate results. When I shared Cobb's (2007) claim that our traditional curriculum is deceptive, Scott reported,

> Right, right… But if you are teaching [the two methods] side by side, you are not totally deceiving students if you come clean and say that this is a decent approximation of the binomial distribution.

Simulation as a statistical method. As far as using binomial simulation as a statistical procedure, Scott correctly noticed that "the results [of the simulation] approach the exact binomial probabilities as we increase number of trials." This indicates his understanding that the simulation is more precise than the normal distribution approximation. Therefore, as far as the precision of the results, with sufficiently many repetitions, Scott reported favoring the binomial simulation.

Randomization Test

Exploring the simulation. After reading the randomization example about comparing mean blood pressure reduction levels of individuals following regular and fish oil diets, Scott noted that the fish oil diet appeared to be more effective, however, he correctly suggested that the difference in means of 7.714 might not be large enough given

the small sample sizes. When Scott initially observed on the screen the reassignment of participants into different groups, he quickly understood the idea of randomization, without much guidance, as demonstrated by the interview excerpt below.

> Researcher: Let's run the simulation [performs the first reassignment of participants into the two groups]. What do you think is happening?
>
> Scott: Are we reassigning the same people into different groups?
>
> Researcher: Yes
>
> Scott: Maybe not all the folks from one group to another, just totally randomly.
>
> Researcher: Exactly. Do you understand why?
>
> Scott: [paused for about 5 seconds] Oh, because if the fish and the regular had nothing to do with it… OK… then it does not matter whether you put them in the fish or in the regular.
>
> Researcher: Right. The assumption here is that the two means are equal.
>
> Scott: Yes, this is always an assumption of the hypothesis testing, that they are not different. At first, I was like, how did you get new data on those diets if they tried fish instead of regular? But I see that it does not matter!

Scott quickly understood that we were trying to figure out the likelihood of obtaining the difference of means equal to -7.714 or more extreme and was able to complete the problem. He was curious to compare the results of the simulation to those of the *t*-test and found a difference of only .003 in the tail proportions, which Scott viewed as a reasonable discrepancy.

Reactions to the simulation. The instructor realized that the logic of the randomization test was different from the logic used in the *t*-test for comparing two independent means, saying

> Wait, when you are comparing two means, it's totally different… this is neat, this is really neat. I don't think I would think to shuffle data like that. It think it's cool… but if your null hypothesis is that there is no difference in the groups then why not? You would get the same data if they [the means] are the same.

Scott appreciated the randomization test can also be done for statistics other than the mean and reported, "what if we are comparing incomes? Then medians would be more appropriate." During the interview, several times the instructor indicated that the randomization simulation was "neat," "awesome," and noted multiple times that this approach was "completely different." Scott liked that the method used "the real imperfect data instead of using results based on mathematical ideal." He meant that the simulation did not have a normal distribution assumption, but instead the distribution was generated empirically based on the given data.

Using the simulation for teaching. While Scott seemed fascinated by the logic of randomization test, he did not think that students would benefit as much from this program as from the binomial simulation. Scott reported,

> I think that the first one [simulation] takes the same concepts and makes them seem even clearer. This one [randomization test] is a little different. It's different than most people think of hypothesis test but really neat. I think this is cool for seeing a different perspective. This is definitely a different perspective. The first one seems like it's just adding to the explanation without increasing to overall knowledge base, without trying to shovel more content into the students' brains. This one is a different enough approach that it would be throwing more into their brains, but it is a very cool thing to see.

Scott was concerned that the simulation approach would not help students understand the formal method of hypothesis testing involving two independent means, as it involves different reasoning. For this reason, he did not recommend using this program as much as the binomial simulation.

> Ah, I don't know if it is as easy to see that this should work out to give you the same or at least very similar results to a traditional 2-sample test…I would use it, but I think I would use it less. That shuffling thing might seem confusing to students…Maybe I would do it with a class or two and go from there.

Scott did not want to make a definite recommendation before trying the program with students. He correctly saw the different logic between randomization and the formal approaches but concluded that using both methods would confuse rather than help students. The instructor was reluctant to recommend giving up the formal methods. One reason is that Scott had a difficult time envisioning students' assessments with simulation-based curriculum. Second, he was uncomfortable with the idea of a statistics course without any formulas, stating, "They are awesome tools to use in a classroom**.** Doing the whole course though? Would most formulas get cuff off entirely? That's strange."

Scott did not want to eliminate the formal approach because the instructor likes connections between mathematics and statistical methods and feels that formal methods provide these connections.

Simulation as a statistical method. Scott was resistant to give up the formal method of comparing two independent means because he believes that "mathematics in statistics is needed." However, Scott did not think there is a flaw in using the

randomization simulation as a statistical method and believes that "even though the fundamental approach is different, they have the same core fundamentals." Scott explained that both methods were making the same assumptions of equal means. However, while the formal method compares two distributions, the randomization test combines them into one distribution of mean differences. It appears that Scott did not have a problem using the simulation as a statistical method instead of the formal approach of testing two independent means because the results were very close, which was not the case for the binomial simulation and its corresponding test.

Instructor Morgan

Instructor Background

Instructor Morgan has been working at CCCC for two years with a total of 23 years teaching at other community colleges and high schools. While she has taught a variety of courses ranging from arithmetic to Calculus II, Morgan started teaching statistics two years ago for the first time. The instructor has both bachelor's and master's degrees in mathematics.

Attitudes toward Statistics

Morgan took one undergraduate and one graduate level statistics course. She had a positive experience as a student and reported enjoying both classes. However, she did not retain much information and had to relearn almost all the material before teaching the course.

Morgan likes the course, although the first time teaching required a lot of preparation. The instructor shared:

> I've been teaching for a long time, but I had never taught statistics, so I had to make sure that I understood it and knew why it was the way it was. So, that was a little bit of work, but I enjoy this class a lot.

Morgan likes that statistical concepts can easily be applied to students' lives and thinks "It's fun for the students. They can relate it to stuff that they see and deal with all the time… That's a lot different than your pre-calculus class, so that makes it really enjoyable."

Morgan finds it most rewarding to teach students "to be critical of what they see," and to be able to use real data for examples, which students often find interesting. For example, she used data from U.S. department of Labor on students' employment rate and year of education and demonstrated that there was a correlation between the two variables.

Morgan's least favorite topic to teach is hypothesis testing. She expressed, "Some of the stuff, I feel they will never do unless they are either doing research or become statisticians themselves. Like hypothesis testing, they are not even gonna do that in the future." For this reason, Morgan believes that it's not necessary for the students to know the computations, but they should have "the actual understanding of what it [the result] means" in order to understand and evaluate research studies.

Perception of Student Understanding of and Difficulties with Statistical Concepts

Morgan believes that students struggle the most with confidence intervals and hypothesis testing. She reported that students typically do not have difficulties with computations and procedures, but they struggle to understand the concepts. However, the latter is exactly what the instructor wants students to get out of the course, as explained in the interview excerpt below:

> Understanding of why they are doing what they are doing and what the result is saying to them in terms of the original question is very challenging to them... Sometimes I think, like in an algebra class, they just want an algorithm that they follow and get an answer, and they aren't necessarily paying attention to the understanding of what this is, what it represents. They just want to get the numbers… [Morgan points to a spot on the paper.] There is my answer… But that's not what I want, I want them to understand what this confidence level represents… And there is a gap between "I know how to do the calculations on my calculator" and the understanding of it.

In summary, it appears that the Morgan strives to provide a foundation for students' conceptual understanding of statistical ideas so that the learners can critically evaluate quantitative information. She does not like teaching inference because most learners are only worried about obtaining the correct answer, making it difficult to foster an understanding of complex inferential concepts.

Binomial Simulation

Exploring the simulation. Morgan was quick to understand the binomial simulation with a coin example and was able to make and verify the predictions. She reasoned, "It's unlikely to have 9 or 10 heads… it's unlikely that this coin is fair."

When we switched to the drug problem, it took Morgan some time to realize what the spinners represented, asking "Is blue success? And the black line is where it is falling, yes?" Once Morgan figured the meaning of spinners, she was able to set up the problem and to correctly solve it.

Reactions to the simulation. In Morgan's view, the binomial simulation, with 50% probability of success, would be easier for students to understand than the drug problem, with the assumed improvement rate of 40%, because the latter problem involved two different drugs. Nevertheless, she liked the simulation and found it useful:

> I like the visuals; I like the simulation a lot. This is the best way to see for them where it is likely, where it is unlikely. For the students that's really effective… It's interesting because we bring the whole idea of the confidence level and where the cut-off is… and show what's reasonable and what's not reasonable so they can make those connections.

Morgan was referring to making connections between confidence intervals and hypothesis testing, which the simulation allows. She believes that, in general, simulations are great tools for visualizing statistical concepts. Morgan uses a similar visual in class; A program that simulates the rolls of two or more dice, then computes the average of *n* rolls and constructs a histogram of sample means. The professor uses the program to discuss with students which values for the average are common and which are rare and to introduce the concepts of critical value and critical region.

Morgan also reported that students often confuse the sample size with the number of times the experiment is performed and thought that the simulation would help them clear this confusion.

Using the simulation for teaching. Morgan reported that she would use the binomial simulation in the classroom as it would help students understand the big picture of hypothesis testing. The instructor views formal methods in an introductory course as unnecessary because the "nitty-gritty calculations" prevent students from understanding important concepts:

> I think simulations are important. Well, in a very introductory course I can see completely throwing out the calculations, because, as I said, they are not really ever going to do that… But understanding is important, like why it is a critical region… it's so abstract when it's just pushing buttons on their calculator. They [students] don't understand what's happening and what is literally taught.

Although Morgan is in favor of abandoning formal inferential methods, she is concerned about testing students with simulations. When I suggested assessing student learning in a computer lab, Morgan was still uncomfortable with them getting different answers for the *p-value* and reaching different conclusions. We then discussed that test results would be more similar if we repeated the experiment many times. After a short pause, the instructor concluded, "Yeah, I think in a very introductory course, just to explain the concepts of what's happening, I can see using only simulations. So, I would say yes [to using only simulations]."

Simulation as a statistical method. Morgan did not expect the difference between exact binomial probabilities and a normal approximation to be so large. She remarked, "So if someone had done a study and this is what they got [pointing at the binomial probability of 0.0565], they would conclude that drug Y is not better, but the

normal approximation shows to reject [the null that they are the same]. I don't know what to say…"

The drug example made Morgan doubt the formal method of testing for one proportion, and she suggested that it was necessary to address this discrepancy in class.

Randomization Test

Exploring the simulation. After reading the blood pressure reduction problem, Morgan noted that the fish oil diet seemed more effective, however the sample size was too small to make a definite conclusion. As we discussed reshuffling the participants into different groups, she asked, "So everyone go to a normal diet or they stay on the diet that they are on?" Morgan was initially unsure what the simulation was doing, asking questions, "So the red ones are all these original people, yes?" or "Are they keeping their scores?" After a minute, the instructor figured out the logic of the test, noting,

> OK, I see what's happening. They are keeping their scores, and I reshuffle them… I got it! We are going to keep doing it [reassigning the participants into groups] so we get the whole bunch of differences. Ok… and we are going to compare and see if this is [the difference obtained in the study] unreasonable.

After this, the instructor finished up the problem and concluded that the fish oil diet is more effective.

Reactions to the simulation. Morgan stated that it was not very easy to understand the reassignment of participants into the regular and fish oil diet but found the process very logical. Morgan reported, "I think the hard part is just recognizing

reshuffling, but we are assuming no difference, so that makes sense. I like that! It's really cool!"

The instructor thought that randomization reasoning would be more understandable to students than the current method of teaching *t test* for two independent means.

> I think that's really cool. I think they [students] would understand this idea actually better than the way we teach it right now. They will get it. I am assuming they [the means] are equal, so it does not matter where they [the participants] will go. I think that they can really understand this simulation. It took me a second… I can reshuffle them because my assumption is that there is no effect… It makes perfect sense that that [-7.714] is an extreme value.

Morgan also noticed that there was no normality assumption for this method, which she found interesting and beneficial. However, she initially did not recognize that the logic of the test is different than that of the formal test of two independent means. We then discussed that in this method the researchers are not generalizing a bigger population as there is no assumption of random sampling. After that, Morgan saw the differences between the formal and informal approaches, saying, "yes, it is a different idea."

Using the simulation for teaching. Morgan found the randomization simulation even more effective than the binomial one. She shared, "Once you start comparing two populations or two sets of data students have a really hard time with it."

Morgan expressed that she would definitely use the program in the future. If the college requires teaching formal methods of inference, then she would start the topic with the simulation so that the students see the reasoning process and then do the "nitty-gritty

with using symbols and formulas." However, given a choice, Morgan would prefer teaching students inference only using simulations.

Simulation as a statistical method. As far as using the randomization simulation as an instructional method, Morgan did not give a definite answer. However, she kept stressing that a simulation-based course would be beneficial for those students "who will not do research in the future." It appears that the instructor views formal methods as the correct way to perform statistical analysis compared to simulations, which are simply tools for explaining concepts to students.

Instructor Cassidy

Instructor Background

Instructor Cassidy has a total of 21 years of teaching experience, which includes teaching at various two-year colleges and high schools. She has been teaching at CCCC for five years and was primarily hired to teach non-college level mathematics. However, in addition to developmental courses, she has taught pre-calculus, linear algebra, and elementary statistics.

Cassidy has a bachelor's degree in mathematics and a master's degree in education. She took an undergraduate statistics course and three graduate-level courses (Statistics for Engineers, Regression I, and Regression II).

Attitudes toward Statistics

Despite teaching elementary statistics for four semesters, Cassidy does not feel fully prepared to teach the course. She expressed, "I am not as comfortable with it

[statistics] as I am in an algebra-based class…I have a little bit of anxiety before I teach it." Cassidy has not been teaching the course continuously, so the instructor needs to review and "re-learn" the material before teaching it every time, after which she gets more comfortable and feels prepared.

Cassidy appreciates that statistics is very applied and often tries to relate its concepts to real-life problems. For example, she used to work as a property underwriter at an insurance company. She often uses the example of computing standard deviations for assessing risk as a class example. Cassidy explained, "I tell students that a large standard deviation is equivalent to instability, whereas small standard deviation means more predictability and consistency."

The instructor likes teaching hypothesis testing the most because of "the structure of the 5-step process," and the fact that the same procedure works for all tests. However, she does not find teaching the introductory chapter enjoyable, which includes many definitions such as types of data, levels of measurement, and others. Cassidy shared, "It's not math. I like calculations and chapter one does not do that."

Perception of Student Understanding of and Difficulties with Statistical Concepts

Cassidy believes that in general students "understand" statistical concepts, provided they spend sufficient time studying. The instructor shared,

> My students this past summer understood everything. Quite a few got A's in the class, so they understood. I don't think students who did not get good grades spent enough time, and that was the reason why. Students who put the time and did the homework assignments seem to really do well in class.

Cassidy thinks that students did well in the summer class because they were required to do all computations by hand. The instructor is against using calculators and believes that students learn better when they work out problems by hand.

> This is the first time I did not use the calculator. You know, I had them use the formulas and stuff and it worked fine. They did fine because they did everything by hand. We practiced in class, and I gave them pretty much similar problems for the assessments, whereas before, they would have to put things into the calculator, press buttons and get the answers. So, I feel they were not really learning anything.

Students were shown how to compute descriptive statistics on a calculator, but they had to work out formulas by hand and use statistical tables, even for estimating *p-values* for *t* and Chi-square distributions.

Cassidy believes that students learn better when they perform computations manually. However, it appears that she equates conceptual understanding with the ability to follow steps, as demonstrated by the interview excerpt below:

> Researcher: Do you think they found it easy to understand the concepts? I don't mean how to perform computations.
>
> Cassidy: Yeah, yeah, I think so. We did enough problems so they just kind of follow that procedure for every one of them. You get a problem with different numbers and the application of how to do them is still same.
>
> Researcher: I see. How about making meaning of their answers? Do you feel students know what a *p-value* means?
>
> Cassidy: Yes, yes, it's the area. They know that if this area is less that the significance level, they need to reject the null hypothesis.

It seems that Cassidy places emphasis on manual computation and working out many problems in class so that students are able to master procedures.

Binomial Simulation

Exploring the simulation. When we started investigating fairness of a coin by tossing it 10 times, Cassidy predicted that if the coin produced 9 heads, it could be a fair coin. However, she had difficulty making a definite conclusion because of the small sample size. After simulating an empirical sampling distribution, Cassidy correctly noted, "it's probably not fair because it's [tail proportion] so low. If you keep performing it over and over again, it's [9 heads] gonna be unlikely. It might happen once or twice, but it's not gonna happen often."

Cassidy was also able to correctly set up the drug problem, although she wished to test it on more patients. After we discussed that it is often difficult to have large samples in real life, Cassidy noted, "Right, this could be a start." She correctly stated that based on a 5% cut-off a 6% chance was likely so that "drug Y may not be more effective."

Cassidy did not have any difficulty with the program and was able to correctly predict and interpret the results.

Reactions to the simulation. Before I had a chance to ask what Cassidy thought of the program, she remarked,

> I think it's a good way to demonstrate this stuff in class, you know. I think it's [the program] very visual, you know. Rather than reading it from the book, you use the simulation to show, you know, the likelihood or certainty of something happening. Students will see what it really means.

Cassidy liked the program because it gave an opportunity to solve the problems and felt the program would benefit students because "they have a natural tendency to use

computers and electronic devices." Therefore, performing and understanding statistical tests with simulations might be easier for them than working out problems by hand.

Using the simulation for teaching. Cassidy advised using the binomial simulation as a supplement to the formal method to compare the results of the two approaches. she did not feel that the formal *one-proportion z-test* should be replaced with the binomial simulation because the latter method provides structure:

> Cassidy: I think it's important to show them [students] the five steps of hypothesis testing. Why not show this [simulation] as an example to relate it?
> Come up with a *p-value* here and compare what you do in class...
>
> Researcher: Why do you think it's important to teach the 5-step method?
>
> Cassidy: It's structure, you know? You go by a structure by means of testing the hypothesis to get a conclusion… Here [in the simulation] you just throw stuff in there… Everybody will be coming up with a different answer… and where is the level of significance here?
>
> Researcher: You can pick it, just like the researchers do in real life.
>
> Cassidy: Right, right.

Cassidy was concerned if students are not given steps to follow, they will be confused with the process and will not be able to perform the procedure. I suggested giving a structure to the simulation methods, asking students to first identify the assumed proportion, then determine the sample size, after which generate the sampling distribution, and finally, compute the likelihood of obtaining results from the experiment. Cassidy felt that this would give students more "structure" so that "everybody will be at

the same place." However, she still felt that working the problems by hand was important for student understanding.

> We still do the same thing with the formal method, but we do it by hand. Here, the program is doing it for you, whereas the students have to do it themselves. They will have to, you know, look at the chart, figure out, and pull the stuff out. I think you understand, when you do things yourself and you have to set it up, you have a better understanding… I am not saying that this [the simulation] is not good. It's just different. It's like a calculator doing fractions for you and you don't understand how to do it.

When I shared the concern that while students are able to follow the procedure for hypothesis testing, they struggle with interpretation of results, such as the meaning of the *p-value*, Cassidy agreed. She stated that that most students probably lack the understanding of this concept and added that they would probably never have to perform calculations by hand. She felt students are likely to understand the program and the meaning of the concepts. In addition, Cassidy remarked that the advantage of using simulations alone is that they would free up time for coverage of other important statistical tests. Yet, the professor still felt that the simulation should be taught in parallel with the formal method.

Simulation as a statistical method. Cassidy noted a difference in results for the binomial simulation and its corresponding approach, but was not as concerned about it. The instructor did not feel qualified to make any comparison between the use of the simulation as a statistical procedure.

Randomization Test

Exploring the simulation. After reading the problem about effectiveness of fish oil diet, Cassidy recognized right away that a two-sample *t-test* would be appropriate for the data. As we started re-assigning participants into groups, she became confused and stated that it was tricky. She thought that people follow a different diet once they are switched to a new group. After I explained the reshuffling process, Cassidy was able to calculate the tail proportion and conclude that the fish oil diet was more effective.

Reactions to the simulation. Cassidy expressed that the simulation provided a different way of thinking about the problem. The instructor expressed, "It's a good way for interpreting the problem. It's a good learning tool. It makes sense that the fish oil is more effective, because the difference between means is so unlikely." However, the instructor was still unsure why we re-randomized the participants, expressing, "I am still confused a little bit about the whole reshuffling, but you cannot go by me because I am not an expert in statistics." It seemed that the instructor did not completely agree that we can simply reassign people into groups without performing the experiment over and over.

Using the simulation for teaching. The lack of Cassidy's confidence in statistical knowledge prevented her from making specific recommendations. Before using the simulation, she wanted to think more about it and practice many different examples.

> The way you explained the whole thing was logical, but I don't feel comfortable enough to actually do it in class, you know… I would have to learn more. This is a new way of presenting stuff, a totally different way. I have a harder time learning this way, than I do with paper and pencil from the notes. I could not explain it as well as you did. I would have to take notes and practice, but this does

> not mean others should not use it. Don't go by me. Students might actually do better with this because they are computer savvy.

Cassidy also inquired if there was any study on which approach was more beneficial for students and expressed interest in seeing research on this topic.

Simulation as a statistical method. Unlike the binomial simulation, which resamples the corresponding formal method closely, the randomization test involves a different reasoning than the formal test of comparing two independent means. While Cassidy understood the main idea of the simulation, she struggled with some parts of it, specifically regrouping data, and did not feel ready to provide any recommendations for using simulations as a teaching and a research tool.

Instructor Clark

Instructor Background

Instructor Clark has been working at CCCC for ten years and has a total of 35 years of experience teaching mathematics and statistics at various two-year colleges. Clark has an undergraduate degree in mathematics and a master's degree in statistics. At CCCC, he typically teaches elementary statistics, pre-calculus, and non-college level mathematics.

Attitudes toward Statistics

Clark likes teaching statistics because it is "more applied than mathematics" and is used in many disciplines. He especially likes teaching hypothesis testing because "it's the same procedure across different tests" and because it is very easy to find applications

for it. There are no topics in introductory statistics course that Clark does not like teaching.

Perception of Student Understanding of and Difficulties with Statistical Concepts

Clark feels that students like statistics because "it does not involve a lot of algebra." He expressed, "Many years ago, statistics was very difficult because computations had to be done manually. Now computers make everything easy for us." Clark thinks that students struggle less with statistics because they can perform all statistical procedures on a graphing calculator. Despite that, he believes that most students still have difficulties with hypothesis testing because they are often confused about which test applies to a given problem. Once students figure out the appropriate test, they can easily follow the steps because the procedure is the same for all hypothesis testing problems.

Clark associates student understanding with their ability to perform computations and follow procedures.

> Researcher: You are saying that students can easily follow steps of the process.
>
> Clark: Yes, easily.
>
> Researcher: Do you think they understand the meaning of the concepts. For example, what does a *p-value* mean?
>
> Clark: Sometimes it's hard for them to understand, but most of the time the calculator will compute it.
>
> Researcher: But, do you think they can interpret what it means?

> Clark: Yes, because we tell them if *p-value* is less than or equal to alpha, they have to reject the null. Sometimes they find it difficult, but most of the times it is okay.

In summary, Clark feels that students prefer an introductory statistics course to other mathematics courses because they don't need to have a strong mathematical background to be able to get a good grade in the statistics. With the help of a calculator, learners are able to perform computations, provided they can identify what type of test is needed for the given data. It appears that Clark does not stress interpretation of concepts and is more concerned with students' mastery of computations and procedures.

Binomial Simulation

Exploring the simulation. Clark found the coin problem easy to understand and was able to explain the process. He correctly set up the drug problem and noted that drug Y would appear better than X if it produced 13, rather than 12 improvements in 20 patients. This statement was correct as 13 improvements would result in the tail proportion of about 2%, compared to approximately 6% in case of the 12 successes.

Reactions to the Simulation. When I inquired about Clark's views on the simulation, the instructor reported that it was "nice, because it is different" and that "students might like it." However, he did not seem very impressed with the simulation. "It's a good program, but technology is going to keep improving, after couple of years this might be trash," remarked Clark with a smile. It appears that he is skeptical about technology as it continuously changes and improves.

When we compared the exact binomial probability to a normal approximation, Clark noticed the difference of more than 2.5% in the tail probabilities and remarked that the two methods would lead to different conclusions based on the 5% significance level. However, Clark was not too surprised by this difference and reported that "statistics is based on approximations." He did not expect to see a small discrepancy between exact binomial probabilities and normal approximation as the instructor regarded the formal test as a rough approximation to the binomial probabilities.

Using the simulation for teaching. Clark did not see a need for using the simulation in class because "only about 30% of students are good students who will learn with either method. The rest of the students just want to pass the course and get out of there [the class]." Clearly, Clark thinks motivated students will learn no matter how the material is presented. He does not feel that the simulation would help students understand the material any better than traditional methods. Clark believes that students learn with practice, and "only when they do it on their own." In other words, whether the teacher presents a simulation or a formal method does not make a difference in students' learning, unless they "practice procedures on their own."

As the instructor kept stressing students' practice of procedures, it appears that Clark, again, was referring to mastery of skills and not the understanding of the logic of inference.

When I shared that some institutions were only using simulations for teaching inference, Clark was surprised and could not envision the instruction and assessment of

such curriculum. When I explained that classes were held in a computer lab, and tests were done through the simulations, he expressed concern that this would "reduce students' brain power." Clark feels that with the increased usage of computers and calculators for instruction in the recent years, expectations for student learning has been decreasing. "Before using calculators, everyone knew mathematics. Now they [students] cannot even multiply and divide."

Clark's views are similar to Cassidy's ideas in that both instructors believe that reliance on technology decreases student learning as calculators and simulation perform computations for students.

Simulation as a statistical method. For Clark, the choice of a statistical method did not matter much, because "statistics is an approximation." In other words, in either method, we are not calculating the exact probability. We discussed that with the increase of the number of trials, simulated distribution was approaching the exact binomial distribution. Therefore, the simulation can be more precise than the normal approximation. Clark then suggested that "with health- and safety-related problems, we should use the exact [binomial], otherwise, it does not matter."

Randomization Test

Exploring the simulation. Clark initially had a hard time understanding the reshuffling process, but he quickly figured out the logic, "Aaah, we are assuming equal means so that we can switch people to groups." Clark requested to compare the medians and ranges instead, and got interested in other simulations. We ended up looking at

simulations that compared multiple proportions and means. After which he requested the link to the programs.

Reactions to the simulation. Clark seemed more interested in the randomization simulation than in the binomial program, repeating "nice," and "interesting" several times. Clark was able to correctly predict that the distribution of differences would be normal and to calculate the appropriate probability.

The instructor appreciated that the simulation could test for the difference in medians and other statistics and the absence of test requirements, expressing, "In real life no way you can have a random sample of patients." However, he did not trust this program completely because he did not see why it would produce similar results as the *t-test* for two independent means. Although the results of formal and informal tests for the blood pressure study were similar, Clark wanted to test the program on other data in order to compare the results of the *t-test* and the simulation for different data-sets, stating, "We have to do many other comparisons for other data. It's not clear how it works."

Using the simulation for teaching. Despite of Clark's interest in the program, he was unsure whether students would understand the simulation. He shared, "I like it, but students might not be able to understand this. You have to test it [on students] first."

Simulation as a statistical method. Clark could not provide any recommendations on using the simulation as a statistical tool because he was not sure how close it is to the corresponding formal method. It is interesting, however, that even though Clark suggested that statistics was an approximation, he still viewed the formal

method as the ideal. Clark stated that it was important to test the randomization test on many different data sets and compare it to the formal *t*-test. This statement suggests that he will only accept the randomization method if it provides similar enough results to the corresponding formal approach, which implies that for Clark, the formal method is "the true" procedure.

Several times after my interview with Clark, I inquired again about the instructor's views on the randomization simulation, but he did not have time to further explore the program.

Instructor Cross Case Analysis

Table 6 presents a cross case matrix comparing the results across the four instructor cases in order to answer the following research questions:

1) How do instructors describe their students' understandings and misunderstandings of inference?
2) What are instructors' attitudes toward simulation-based inference?
    i. How do instructors compare formal and informal inferential methods as instructional approaches in introductory statistics courses?
    ii. How do instructors compare formal and informal inferential methods as statistical methods?

The instructors in the matrix are presented in order from most to least open to using SBI.

Table 6

Instructor Cross-Case Matrix

	Morgan	**Scott**	**Cassidy**	**Clark**
Student Understanding of and struggles with Inference	Students can generally follow procedures but have difficulties with language and interpretation of concepts.	Students typically are able to perform computations, but do not understand meanings of concepts.	Most students do well because they have sufficient practice opportunities and are required to perform computations manually. Understanding and the ability to compute are synonymous.	Most students do well because calculator does computations for them. Once students know which test applies, they can easily follow steps because they are the same for every test. Understanding and the ability to compute are synonymous.
Binomial Simulation as Instructional Method	Use the simulation to help students understand the fundamental concepts of inference and make connections between hypothesis testing and confidence intervals.	Use the simulation before introducing formal methods of inference to explain the big picture and go back to the simulation to compare results.	Both simulations are beneficial for computer savvy students. Formal methods are important because they provide "structure" and because students learn better if they work out problems manually.	Students might benefit from the simulation. However, instructional method does not make a difference. Students who are motivated will learn either way.
Randomization Test as Instructional Method	Use it more than the binomial simulation because students struggle more with multiple samples.	Use the randomization simulation less than the binomial one as the connection to the formal test might be unclear.		Might be confusing for students.
Simulations as Statistical Tools	The binomial simulation is more accurate than the corresponding formal method. Unsure about the precision of the randomization program.	Both simulations are good statistical tools. The binomial simulation provides a more accurate result than the normal approximation. The randomization simulation and the 2-simple *t-test* provide similar results.	Doesn't feel qualified to provide any recommendations.	Use the binomial simulation for problems involving "risk." Otherwise, it does not matter as statistics always involves approximations. More comparison between the methods comparing two means is needed.

Research Question 1. How do instructors describe their students' understandings and misunderstandings of inference?

Analysis of the cross-case matrix yielded three assertions described below.

Assertion 1. Students in an elementary statistic course are typically able to perform necessary procedures and computations.

All instructors reported that, in general, their students can follow steps and perform the computations needed in an introductory statistics course because it does not require strong algebra skills. All instructors except Cassidy allow students to perform computations with a graphing calculator. Although Cassidy requires manual calculations, she does not feel that students find them difficult, as they practice many problems in class.

It is not surprising that introductory statistics students find statistical procedures easy as the course only requires basic computations compared to other college-level mathematics courses. In addition, as Cassidy and Clark reported, while different formulas are appropriate for different statistical tests, the general steps are the same, which makes it easy for students to follow statistical procedures.

Assertion 2. Instructors have varied beliefs about what is important for students to learn from an introductory statistics course.

Although all four instructors like teaching statistics because of its wide applicability compared to other mathematics courses, they have different expectations in terms of student learning.

In Morgan's view, most students taking an elementary statistics course will never have to conduct their own research studies. However, they must understand claims made by others and, therefore, should learn how to think critically about statistical information. Thus, Morgan is less concerned with the mathematics behind statistical analysis and places more emphasis on interpretation of statistical concepts.

Scott also values students' conceptual understanding, but unlike Morgan, he stresses connections between mathematics and statistics. Scott derives some formulas using calculus and feels that student understanding of mathematical formulas is crucial.

For Clark, it's important that students understand what procedure applies for a given situation and that they can perform that procedure. He does not stress the reason behind each step of the process and is more concerned with students' ability to decide which test to use and to correctly work it out.

Cassidy appreciates structure in hypothesis testing and the fact that the process is the same across different tests. Just like Clark, the instructor believes that students understand hypothesis testing if they can work out examples. However, unlike Clark, Cassidy requires all computations to be done by hand as she believes that manual computation improves student understanding. Both Cassidy and Clark associate conceptual understanding with procedural fluency.

Assertion 3. While most instructors view inference as the most problematic area, they report different reasons for student difficulties with this topic, including inability to (a) select an appropriate test for a problem, (b) understand meanings of the concepts, and (c) correctly interpret results.

All four instructors reported that hypothesis testing is most problematic for students, but for different reasons.

Morgan stated that students struggle with hypothesis testing because they have a hard time understanding the meaning of the concepts. Therefore, they have difficulties interpreting their results.

Likewise, Scott feels that students typically have difficulties with language in statistics and find the wording of hypothesis testing challenging. In addition, they often get confused whether the test involves means or proportions, one sample or two samples, dependent or independent samples, etc.

Clark believes that since students are allowed to use calculators, they do not struggle as much with statistics. However, like Scott, Clark also feels that students are often confused about which statistical test they should use for a given problem.

Cassidy is the only instructor who feels that students generally do not have difficulties with this topic, provided they put in sufficient time to practice the process. As mentioned earlier, the instructor was referring to mastering the procedure rather than understanding the logic and concepts of hypothesis testing.

Research Question 2. What are instructors' attitudes toward SBI?

Three assertions were developed regarding the instructors' attitudes toward SBI and are presented below.

Assertion 1: All four instructors expressed mostly positive views of SBI, citing various benefits of the simulations.

While all instructors conveyed generally positive attitudes toward simulations, their reasons for this attitude varied.

In Morgan's view, simulations would help students visualize complex statistical concepts and help them understand the big picture of hypothesis testing. She also felt that the programs would assist students in making connections between confidence intervals and hypothesis testing as well as in understanding the difference between sample size and number of samples. Morgan found randomization test more beneficial than the binomial simulation, because students typically struggle more with statistical inference involving two samples.

Like Morgan, Scott also commented that both simulations were very visual and are great tools for introducing the topic to students. He also appreciated that the binomial simulation clearly depicted that the sampling distribution is discrete. Since the formal method approximates a discrete binomial distribution with a continuous normal curve, students often incorrectly assume that the binomial distribution is continuous. Scott also liked that the randomization test allows for comparison of medians and other statistics, whereas the corresponding formal methods is performed for only means. Finally, he

reported liking the simulations because they deal with real data instead of being based on "ideal" mathematical models.

Cassidy thought the simulations would benefit students since they are generally very good with technology and therefore, might find a computer program easier than working out problems by hand.

Clark reported liking the binomial simulation, but seemed indifferent toward the program. He reported that with the constant change in technology, the program might not be useful in the next few years. Clark expressed more interest in the randomization test and appreciated that this method did not require a normal assumption and random sampling. Similar to Scott, Clark also liked that the randomization test can be performed for statistics other than the mean.

All instructors, except Scott, found the re-shuffling process slightly unclear at first, but understood it quickly. The only exception was Cassidy, who seemed to understand the logic of the process, but was not quite sure why re-assignment of participants into groups was possible.

Assertion 2: Most instructors regard the binomial simulation as a better statistical approach than its corresponding formal test. They view the randomization test as either comparable to its corresponding formal 2-sample *t*-test, or have no opinion on this matter.

Scott correctly indicated that the binomial simulation is more precise than its corresponding formal method when the number of trials is sufficiently large. All instructors except Clark did not expect to observe a large difference in tail proportions for

the two methods (close to 3%) and felt that it was important to address this discrepancy in class. According to Clark, the binomial simulation is more appropriate for "serious problems" that are concerned with health and safety, but the choice of the method does not matter for less important issues as statistical procedures are not precise in general. Cassidy did not feel qualified to make any recommendations regarding the usefulness and precision of either method.

As far as the randomization test, the instructors were less comfortable making a comparison between the simulation-based and formal methods. All of them appreciated that the simulation works for statistics other than the means, and that it does not have a random sample and normal distribution requirements. However, Scott was the only instructor who thought that simulation-based and formal methods in general would produce similar results. The other three instructors did not feel comfortable enough to comment on the effectiveness of the simulation as a statistical tool. Clark expressed interest in testing the simulation on different data sets and comparing the results to the formal method, but did not have a chance to do so. Nevertheless, he doubted the randomization simulation and would only accept it as a valid statistical method if it was close enough to the *t*-test.

Similarly, Morgan seemed to perceive formal methods as more mathematically correct. She expressed several times that SBI would be useful in a very introductory statistics course, implicitly indicating that formal methods still should be used in more advanced courses.

Assertion 3: Instructors provided different recommendation for using SBI in the classroom ranging from using simulations as a substitute for traditional methods to not using them at all.

Table 7 provides a summary of the instructor recommendations for teaching statistical inference.

Table 7

Instructors' Preferred Teaching Approach

	Morgan	**Scott**	**Cassidy**	**Clark**
Formal	✗	✓	✓	✓
Binomial	✓	✓	✓	✗
Randomization	✓	Unsure	✓	✗

Morgan was the only instructor who recommended abandoning formal methods of inference in an introductory statistics course because formulas and symbols hinder student understanding of the logic of inference. She believes that since most of the students will never perform computations by hand, they do not need to know the formal procedures. However, they need to be able to interpret statistical concepts. Although Morgan could not envision student assessment via simulations and was concerned that learners would be getting different answers, she still advocated for the replacement of formal inference with SBI.

Both Scott and Cassidy suggested using a combination of formal and informal methods when teaching inference. The instructors expressed that using the simulations would be beneficial for introducing students to the concepts of interference and to

compare the results of formal and informal approaches. Although Scott was surprised by the difference between the outcomes of the binomial simulation and the normal approximation methods, he still did not want to give up the formal method. According to Scott, it's crucial that students understand the mathematics behind statistical procedures. As far as the randomization simulation, Scott felt that the logic of the process is different from its formal method and, therefore, he was worried that presenting both methods to students might hinder their understanding.

Cassidy's reasons for employing both methods were different from Scott's. Because the instructor herself prefers learning using "paper and pencil" and formulas, Cassidy believes that some students would benefit more from the formal methods. In addition, she feels that formal methods provide more structure and, therefore, might be easier for students to follow. Finally, Cassidy is convinced that students learn better from performing computations manually. Since the simulations perform the computations for learners, they limit students' understanding of the process.

Clark was the only instructor who did not see a benefit of using simulations in class. Although he reported liking the programs, Clark expressed that good students learn no matter which instructional approach is used. Moreover, he was not convinced that students would understand the simulations, especially the randomization test, and suggested testing it on students before making any recommendations.

5. STUDENTS' EXPLORATION OF SBI

Lena

Student Background

Lena is a 21-year-old student who is finishing up her associates degree in business administration. She completed a condensed, eight-week summer statistics course three months prior to our interview and received an A with a part-time instructor who was not interviewed for this study. Despite it being a fast-paced class, Lena enjoyed it mainly because of her instructor, who was very organized and detailed. Lena explained, "She [the teacher] had an outline written out for us for every week. Every detail was written out there and all the deadlines. She really broke down every problem."

Student's Prior Knowledge of Hypothesis Testing

Lena felt that at the time of taking the course she had a good understanding of concepts related to statistical inference. She especially liked hypothesis testing because "it was pretty much the same every time and she [the instructor] would tell us why we were doing things."

Although Lena felt she had a good grasp of hypothesis testing, she did not retain much information and only had a partial understanding of the concepts. She described the purpose of hypothesis testing as "to prove whether something… whether statistic is reasonable or not, whether it's within a certain range of probability of happening... Whether it's useful data or not."

Lena was correct in saying that when testing a hypothesis, we examine whether our sample statistic is reasonable (under a certain assumption about the population parameter), but hypothesis testing is certainly not used for determining the usefulness of data.

Lena did not remember much about null and alternative hypotheses other than that "they are opposite of each other." She also could not fully recollect what a *p-value* was and said that it is "the number that helps you decide the usefulness of the test and whether or not the null is true." Although Lena remembered that a *p-value* is used to make a decision about the null hypothesis, she could not recall how.

Binomial Simulation

Student Exploration of the simulation. Lena correctly predicted that the distribution of the number of heads out of ten tosses of a fair coin "would be normally distributed with five at the mean." She explained, "Over the time… over many, many tests that's where the mean would fall. If you do it only couple of times, it's not reliable." Lena meant that with only ten coin flips the number of heads is difficult to predict, but, with the repetition of the experiment, most of the times we would obtain five heads. Lena demonstrated an understanding of the empirical distribution and correctly indicated that "the farther away we move from the center, the less likely [obtaining that number of heads] will be."

I then asked Lena to make a conclusion about the fairness of a coin that produced nine out of ten heads. She correctly indicated, "I would say it's not fair, but there is still a

chance that it could be fair." While her reasoning was correct, she initially struggled to relate her decision to the empirical distribution. Instead, she was trying to calculate the probability of the coin being fair, saying "There is a 10% chance for it to be fair [one out of ten]."

Lena's reasoning demonstrated that she has a common misconception that when testing a hypothesis, we can calculate the probability of the null hypothesis being true (Castro Sotos et al., 2009; delMas et al., 2007). However, computing this probability is impossible. We can only calculate the likelihood of obtaining the results of the experiment, given the truth of the null hypothesis.

When I prompted Lena to calculate the likelihood of tossing nine or ten heads with a fair coin using the simulation, she found the probability to be 1% and correctly reported, "Oh, I see. I would say it's not fair. Normally, with an even coin, there would be a 1% chance that that [result] would occur, so, if it's fair, it is very unlikely that that [9 out of 10 heads] would happen."

I then asked Lena if her conclusion would change if we obtained eight instead of nine heads. She calculated the tail proportion to be 5.5% and reported, "I am not sure, but it's much more likely to happen." As her cut-off between a likely and unlikely result, she chose 10% and correctly reasoned that with eight heads the coin would be considered as unfair (tail proportion = 5.5%), but with seven heads (tail proportion = 17.9%) as possibly fair.

The drug example was a little more challenging for Lena. She correctly predicted that if drug X, which had a 40% improvement rate, was given to different samples of 20 patients, most likely eight patients would improve, with a slight variation. She reported, "40% on average would improve, but could be more or less." While the student had a good understanding of the empirical distribution and correctly indicated which numbers of improved patients for drug X were likely and which were unlikely, she struggled to use the empirical distribution to decide whether drug Y, which improved 12 out of 20 patients, could be considered more effective than drug X. Lena calculated that 12 out of 20 was 60% and decided that drug Y was more effective than drug X because 60% is larger than 40%. Lena failed to recognize that drug Y was only tested on 20 patients and could produce such high success rate by chance. The following interview demonstrates Lena's gradual understanding of the decision process:

> Lena: I would say that drug Y is more effective than drug X.
>
> Researcher: why?
>
> Lena: Because it has a greater chance of being successful.
>
> Researcher: Right, but we only tested drug Y on 20 patients, whereas we know that drug X has the true improvement rate of 40%.
>
> Lena: So you just have to test this one [drug Y] more.
>
> Researcher: That would be nice, but what if we can't? what if it's too expensive? Can you try to make a conclusion with these data?
> [pause]
>
> Researcher: Can drug X improve 12 patients?
>
> Lena: Yes

Researcher: What is the likelihood of that?

[Lena types 12 in the program and finds the tail proportion of 5.9%]

Researcher: What does this mean?

Lena: There is a 5.9% probability that in a long run you would get 12 or more successes.

Researcher: Out of?

Lena: Out of 20.

Researcher. With which drug?

Lena: Drug X.

Researcher: Is it likely?
[pause]

Lena: It's not very likely, but it's definitely possible.

Researcher: What would you conclude?

Lena: I would say it's possible that they [the two drugs] are the same.

Researcher: Because?

Lena: Because drug X could have gotten the same results.

When I inquired whether Lena would change her conclusion if instead drug Y produced an improvement in 13 patients, she quickly entered 13 in the program, found the tail proportion of 1.6% and correctly concluded, "I would say that Y is better than X because it [this result] is unlikely with X."

It appears that Lena struggled the most with the *comparison* stage (stage III) of the conceptual framework. Even though she understood the sampling variability

produced by drug X and knew which results were improbable, she had a hard time using this information for making a decision about drug Y. The guiding questions helped Lena clear this confusion and relate the results of drug Y to the sampling distribution of drug X. Once she made this connection, Lena was able to move to the *explanation* stage (stage IV) and make the correct conclusions for both significant and non-significant results.

Making connections between formal inference and SBI. Initially, Lena struggled to relate the program to the corresponding formal test probably because she had a vague memory of what she learned in the course. She originally thought that for the coin example the null hypothesis was obtaining at least nine out of ten heads. When I explained that the null hypothesis was the assumption we made for the construction of an empirical distribution, she quickly corrected herself, "Oh, the assumption is that you would get heads five times and tails five times, so it [the coin] is fair." She did remember (although she was unsure) that the tail proportion corresponded to a *p-value*, and that "a low *p-value* means that the null is false." In addition, she recalled, "we used [in class] anything below 5% as a low chance."

Lena was able to correctly formulate the hypotheses for the drug problem and reason, "They can't conclude that Y is better than X because the chance of having 12 successes or more would be possible with drug X. So, you won't reject the null that Y and X are the same."

Student's view of the simulation. Lena liked the simulation because "it's very simple looking and not complicated." She elaborated,

> I think once you realize and understand the basic aspects of it, like where the number of samples and sample size goes, then it should be very easy to use… I can see, what's reasonable and what's not by looking at the graph, and I can see the *p-value*.

Lena liked that the simulation provided a clear visual for the problem. She expressed that the program was more useful than a calculator because while a calculator provided a graph of the standard normal distribution with a shaded area, the simulation "shows the graph as it's happening." In other words, Lena liked to observe the process of constructing the sampling distribution, rather than just seeing the end result.

Randomization Test

Student Exploration of the simulation. After reading the problem about fish and regular diets, Lena correctly predicted, "Probably the fish oil diet is better because most of the participants have their blood pressure going down."

While Lena needed some assistance in understanding the reason for re-randomization, she quickly explained, "so we want to see what are the possible differences in results. If the diets are the same, we can move numbers from one group to another."

Lena seemed to understand the empirical distribution of the re-randomized differences. She said, "So, if the fish oil diet and regular oil diet had the same effectiveness, the differences in the means should be most likely zero." She correctly explained that the likely differences in mean were "between -6 and 6, and outside this range they [the differences] are unlikely."

Lena also correctly stated that at the next stage we had to compare the result of the experiment to the empirical distribution and calculated the tail proportion of 0.0008. While she had the correct interpretation of the tail proportion and knew that it was very small, she still concluded that the two diets were equally effective. She reasoned, "If both groups have the same effectiveness, then the difference should fall on that curve… It [the observed difference of -7.714] is not very likely, but it falls within the curve of no effectiveness." Lena argued that even though our result is unusual, it can still occur if the two diets are equally effective. Therefore, she did not feel comfortable with rejecting the null hypothesis of equal means.

Reluctance to reject the null hypothesis even with a small *p-value* is documented in the literature. Liu and Thompson (2005) reported that teachers often demonstrate "commitment to null hypothesis." They reason that even though sample results are unlikely, they could still occur.

Once I reminded Lena that just like with previous example, at some point we need to reject the null hypothesis, she reported, "So there is a 0.8% chance that there isn't any difference… The chance for it helping is much greater. So, the fish oil diet is a more successful treatment."

While Lena correctly concluded that the fish oil is more effective for reducing blood pressure, she misinterpreted the *p-value* as the probability of the null hypotheses being true. It appears that her misconception was so deep that it was not changed by the

program in such a short time. We then discussed again that a *p-value* is a conditional probability and is calculated under the assumption of the null hypothesis being true.

Student's view of the simulation. In Lena's view, the randomization test is more difficult to understand than the binomial simulation, but "it's still a great visual and is definitely useful." The student thought that the interpretation of the differences and knowing what they represented was not easy and the program "definitely needs explanation with it, it cannot just be [used] by itself." While Lena understood the re-shuffling process, she reported that the process was a bit confusing and "needs assistance or guidance, like a teacher." Nevertheless, she liked the dynamic representation of reshuffling and that the program clearly depicted how the graph [of the shuffled differences] was constructed. Although Lena found the randomization test more complex than the binomial simulation, she still felt that the program would "only benefit, I don't think it would hinder learning."

Student's Learning Preference

Lena viewed both simulations helpful for understanding and visualizing the concepts of hypothesis testing and recommended using them together with the formal methods. She clarified,

> I like both [methods] because I think that students learn differently. I am a visual learner and I don't like numbers, but I think both are beneficial, because this [simulation]would be a good visual, but maybe people who are better in math will like to have formulas written out… Sometimes formulas do help me too. I need as many sensory aids, such as hearing, seeing, touching or manipulating … Maybe use the formal methods first and then have these [simulations] to reinforce it. Like, if you didn't understand one way, you could go though it the other way.

In summary, Lena felt that using both formal and informal methods would help students understand and reinforce inferential concepts. She expressed that it is beneficial to have different representations, such as hearing the teacher's explanation, seeing visuals, writing formulas, and manipulating paraments within the simulations.

Natalie

Student Background

Natalie is a 19-year old sophomore planning to major in psychology. She completed elementary statistics, taught by instructor Clark, six months prior to our interview. Although the class heavily relied on a calculator, Natalie found it very difficult and reported that her "main struggle was with equations and memorizing what goes where." Despite Natalie's belief that she was "weak in math," she worked hard and "managed to get a B."

Student's Prior Knowledge of Hypothesis Testing

Natalie did not remember much about hypothesis testing. She reported that its purpose is "to figure out if the data is correct." Although she remembered the terms associated with the test, such as null and alternative hypotheses, significance value, and *p-value*, she could not recall what any of them meant. She reported that while taking the class, she knew "how to do stuff" but did not understand the meaning of ideas. She explained, "With math I am always like this. These are the steps; this is how you do it."

Binomial Simulation

Student Exploration of the simulation. Natalie was able to correctly predict that with ten flips of a fair coin, the distribution of the number of heads resembles "a bell curve with five in the middle." She correctly interpreted the empirical graph and indicated that getting the number of heads close to five was likely. However, as we move away from the center, the probability of getting that number decreased. Although Natalie seemed comfortable with screen representations and sampling distribution, she struggled at the *comparison* stage as illustrated below:

> Researcher: Imagine you have a coin, and you don't know whether or not it is fair. You flipped it 10 times and obtained nine heads. What would you conclude?
>
> Natalie: It [the coin] would definitely favor heads
>
> Researcher: why?
>
> Natalie: because it landed on it nine heads out of ten flips.
>
> Researcher: Can this graph [distribution for a fair coin] help you to decide?
>
> Natalie: Since it's unfair, no. Because it's unfair, right?
>
> Researcher: We don't know, that's what we're trying to figure out.
>
> Natalie: We don't know, right.

At first, Natalie did not see the connection between the result of the experiment and the empirical sampling distribution because she was convinced that the coin was unfair and, therefore, viewed the graph for a fair coin as being irrelevant. Her other issue was that she was trying to draw a definite conclusion about fairness of the coin. Even

with seven heads (tail proportion of about 18%) she concluded, “It’s more likely unfair than fair.”

Natalie was trying to form her conclusion in terms of “unfair” versus “fair” rather than “appears unfair” versus “possibly fair.” When I clarified that we were not trying to prove that the coin was fair, her reasoning changed, and she began incorporating probabilistic language in her explanations. Natalie correctly concluded that for seven or eight heads, “It could happen, the coin could be fair,” but for nine heads, “It could happen with a fair coin, but odds are, it wouldn’t.”

Natalie seemed to understand the logic of inference. She had less trouble with language for the drug example. She correctly interpreted the graph for the number of improved patients with drug X, stating, “So, *x* [axis] is the number of successful patients that improved out of 20 and the *y* [axis] is the number of times that improvement happened.”

However, when I asked if drug Y was better than X given that it worked for 12 out of 20 patients, Natalie concluded, “I would say Y is better, just because 10 is 50% and we have 12, so that means we got more than 50%.” She failed to account for sampling variability when drug X is administered to different sets of 20 patients. I then inquired if the same result could happen with drug X, after which she quickly realized that it was possible. She calculated the tail proportion to be about 5.5% and concluded, “Yes, it could, but 12 is out here in the tail. It’s not likely that it would happen [with drug X].” Natalie also correctly reasoned, “With 13 improvements the probability is 2%,

which is even less likely... It's unlikely for drug X to reach 13 improvements, so I would say that Y is better." However, if less than 12 patients improved with drug Y, she concluded, "It's more likely to get this [result] with [drug] X, so we cannot conclude that Y is better [than X]."

To summarize, even though Natalie needed some guidance at the *comparison* stage to relate the drug Y result to the sampling distribution for drug X, she was able to reason correctly. She considered 10% as her cut-off value between usual and unusual results, and based her conclusions accordingly. She also properly phrased a non-significant finding in terms of "they could be the same" rather than making a deterministic conclusion, such as, "they are the same."

Making connections between formal inference and SBI. Natalie was unable to relate the presented problems to the formal inference possibly because, as she stated, she did not have a good understanding of the process in class. She remembered, though, "If it's [test statistic?] in the tails, it's unlikely, and if it's in the middle between the certain range than it's likely." Natalie did not have a recollection of the null and alternative hypotheses. However, she remembered, "If *p* is low, you let it go." Natalie laughed when she said this phrase and confessed that it she never understood its meaning in class. However, now she had a correct interpretation of the *p-value*. The student reported, "This [the simulation] helps me understand that the 2% [the tail proportion] means that 2% of the times it happening that 13 or more out of 20 patients would improve, so I understand what it means."

When Natalie tried making connections between the binomial simulation and hypothesis testing, she initially got confused and thought that with a low tail proportion she was rejecting the claim that drug Y is better than X. It seemed that when Natalie tried recalling the rule for rejecting the null hypothesis and what the null represented, she stopped thinking logically and tried to follow the process she used in her statistics class. Natalie said, "So when *p* is low, let it go, so you reject your hypothesis that Y is better than X. Isn't it what we are doing?" After about half a minute of thinking about it, she correctly concluded, "No, if the percentage is low, you reject that X and Y are the same."

Student's view of the simulation. Natalie reported that the binomial simulation was easy to understand, and that "the most confusing part was to remember what I learned in stats." She, elaborated,

> I think that if I was to start on this program I could get a lot out of it, especially because it has lots of visual aids in it. That helps a lot. And it's showing you where it exactly is and marking the percentage for you.

Natalie appreciated that the visuals made it easy to understand what the tail proportion represented and made the reasons behind the process of hypothesis testing clear.

Student's learning preference. Natalie shared that given a choice of learning inference with either the formal or simulation-based methods, she would prefer using the simulation approach. She explained,

> This [the binomial simulation] makes a lot more sense that that one [the formula-based method]. In my stats class I kind of tried to figure it out, but it was a little hard. This [program] has a visual aid that helps a lot, so I would prefer this type of class.

When I suggested an option of using both simulation-based and formal methods, Natalie chose using only SBI because "if you try to correspond formulas, it gets confusing." As she shared earlier, the struggled with using formulas in class mainly because she was confused with the meaning of the different symbols. With the binomial simulation, instead of having to memorize notation, she could "look at it [the graph] and figure out what is what and why."

Randomization Test

Exploring the simulation. Natalie found the randomization test less intuitive than the binomial simulation. She correctly interpreted the graph of blood pressure reductions of people in the fish and regular oil diets and predicted, "The fish oil seems better." However, she struggled to understand the reasoning behind re-randomization. She was unsure why we could assume that the two diets are equally effective, asking, "how are you supposed to know if they are the same of not? You did not redo it [the experiment]." It took Natalie some time to understand that we did not know for sure whether the effectiveness of the two diets was the same or different. However, we wanted to obtain possible differences in means assuming equally effective diets. Natalie still thought that moving people from one group to another was wrong without repeating the experiment. She inquired, "When you are putting them into different groups, then it changes the outcome, right?" Again, while Natalie agreed that if the two diets were the same, they would produce the same result, she did not feel comfortable assuming equal effectiveness of the diets.

Natalie correctly interpreted the distribution of the shuffled differences, "The *x* axis represents the differences in the means and the *y* is the amount of times that you shuffled that." She also concluded based on the empirical distribution that having a mean difference of "zero is likely and anything farther than 5 or 6 is unlikely." When I asked her to use the graph for making a conclusion, Natalie located the observed difference of -7.714 on the graph, calculated the tail proportion of *0.07%*, and concluded that based on a very small tail proportion, the fish oil diet is more effective than the regular oil diet. She explained,

> The graph represents that if both fish and regular were the same, these differences would apply. There is 0.0007 possibility to get -7 if they were the same. Because our difference is so unlikely to happen, it's proving that the fish oil is better.

In summary, Natalie understood all stages of the process, but struggled with the reasoning we used for constructing the empirical distribution of mean differences.

Student's view of the simulation. Natalie expressed that the randomization simulation was more challenging "because of the switching." She felt it would be easier if the empirical distribution was provided to her without her having to see the re-randomization process.

> I feel like if they just gave you the graph, it would be more helpful. The graph and having to put the 7 in there made sense. I did not feel that the steps in between, like switching them around was necessary. I kept thinking that they were redoing the experiment, but they were just placing the numbers on different rows and that's all it was.

Natalie shared that while in the binomial simulation the visual aid helped her, in this program it confused her. It seems like she found it counterintuitive to regroup data without repeating the study.

Student's learning preference. In Natalie's view, it would be more beneficial to see both formal and simulation-based methods for comparing two independent means. She found the re-grouping part of the process confusing and for this reason recommended a blended approach.

> It would help me for this [problem] if I got to use the equations and this [simulation] the same time. In the classroom, I felt like I got the gist of it a little faster, and I wasn't as confused. If I were to calculate it myself, I would have a better understanding of why the numbers were there and what they are for. So I think the two combined would work really well.

Casey

Student Background

Casey is a 20-year-old nursing student who completed an elementary statistics class six months prior to the interview with instructor Morgan and earned a C. She took the class as a pre-requisite for a BSN program. Casey explained that toward the end of the semester her attendance and commitment to the class dropped due to family circumstances, and for this reason she ended up with a low grade. She felt the class was not very hard and said, "If you paid attention and did your homework, you could do it."

Student's Prior Knowledge of Hypothesis Testing

Casey reported that she struggled the most with hypothesis testing because she could not differentiate various tests. She felt that she had a good grasp of the concepts, but since procedures are very similar to each other, she often mixed them up.

Casey described the hypothesis testing as an "if then statement" but could not explain what she meant so I provided an example shown below:

> Researcher: Suppose you want to test a claim that the average salary in the U.S. is more than $40,000. How would you go about that?
>
> Casey: I would formulate hypotheses and then gather people. I don't know if you would exactly do the outliers… It's hard to recall.
>
> Researcher: You mentioned hypotheses. What are they and why do we need them?
>
> Casey: I remember that's what you test for and that's kind of your goal for what you want. If you say $40,000 is the average income, then… I am not sure what for then [laughing].

By "if then statement," Casey probably meant that if we assume the null is true, then we can calculate the likelihood of our sample results. While Casey vaguely remembered the idea of hypothesis testing, she lacked understanding of the concepts. She could not recall anything about a *p-value* other than "I do remember that it was very hard."

Binomial Simulation

Exploring the simulation. Casey correctly reasoned that, for a fair coin, out of ten tosses we expect the number of heads to "stay around five, and it's more likely to be more in the middle than below 3 or above 7."

Casey concluded that a coin that produced nine out of ten heads would be "loaded on one side because you got closer to the end of the graph than the middle." She compared the distribution of heads for a fair coin and the given result of the experiment and understood that nine out of ten heads is an unlikely outcome for a fair coin. However, she struggled to draw a conclusion for nonsignificant results. For seven heads (tail proportion of 18.2%), she reasoned, "It's more unlikely that it's likely, so it still favors heads." For Casey, the cutoff percentage between likely and unlikely results was "50%, right at 5, or maybe 6." She explained, "The definition of fair is five [heads] out of ten, so theoretically speaking, it would be hard to say that six out of ten is still fair."

Although Casey correctly interpreted the sampling distribution, she did not consider it when arriving at her conclusion. She failed to take into account the fluctuation in the number of heads produced by a fair coin between experiments. Therefore, she concluded that if an experiment did not produce exactly five heads, the coin had to be unfair. We then discussed that even for a fair coin different results were possible and distinguished theoretical and empirical probability. After that her reasoning improved. In the case of nine or ten heads she concluded, "The coin seems to favor heads." However, with five to eight heads, she inferred, "It could be a fair coin." Casey began incorporating uncertainty in her conclusions, demonstrating an understanding that the result is possible with a fair coin, so it is difficult to make a deterministic conclusion.

Unfortunately, Casey's improvement in reasoning did not transfer to the drug example. She again did not account for variability in drug X results, saying, "Drug Y is

better than drug X" because twelve improved patients out of 20 is greater than 40%. I then encouraged her to use the simulation to examine a possible number of improvements for drug X. Casey correctly set up the simulation and obtained the distribution of the number of improved patients. She noted, "Likely outcomes are between six and ten; others are less likely." She also realized that the peak of the graph was at eight, because it was 40% of 20. However, she had a hard time understanding what the *x* and *y* axes represented. She interpreted, "This [pointing at the *x*-axis] would be the number of trials and this [pointing at the *y*-axis] would be the number of patients, I guess."

We reset the program and started constructing an empirical distribution from the beginning, which cleared Casey's confusion.

> I understand now, I am just trying to put it into words. It [*x*-axis] would be the number of effective trials of one sample size, which would be 20. So how many people you got that were effected positively by the drug. This [the *y*-axis] is the total number you got in that column.

I then prompted Casey to look at drug Y results and form her conclusion. While she was correct in saying that "12 is toward the tail, so probably Y is better," she was unable to explain the details of the process. When I inquired why we used the distribution for drug X to make a conclusion about the effectiveness of drug Y, she answered, "Because it was tested on more people, I guess." I then scaffolded her reasoning as demonstrated below.

Researcher: Could 12 out of 20 patients improve with drug X?

Casey: Yes

Researcher: Is it likely?

Casey: No, but it could still happen [tail proportion of 5.3%]

Researcher: So what would you conclude?

Casey: They could be the same.

With 13 improved patients for drug Y, Casey concluded, “They can still be the same, but it’s just very unlikely. I would probably say that drug Y is better.” While her conclusion was correct, her explanation suggests that she incorrectly interprets the tail proportion as the likelihood of the two drugs being equally effective.

In summary, Casey struggled at all stages of the inferential reasoning process but needed the most assistance at the *comparison* stage.

Making connections between formal inference and SBI. Casey recalled that the tail proportion is the *p-value*. She remembered, “If it [*p-value*] was less than a certain number you don’t exactly throw it out but it’s very unlikely.” She did not quite remember what she was rejecting, calling it “the claim,” and had no recollection of null and alternative hypotheses. She also did not remember the cut-off value between likely and unlikely, thinking it was 50%.

Student’s view of the simulation. Casey reported liking the program overall, but she found it a little confusing at first. She shared:

> “There is a lot of numbers here [pointing at sample size and number of repetitions]. I think if I use it more, then I can probably figure it out, but it’s new to me, so it’s just difficult to visualize just because it’s new. “

Casey struggled at the *recognition* stage of the process. She had difficulty interpreting the graph and seeing the difference between sample size and number of samples.

Student's learning preference. Casey described herself as "old school" because she likes "pen and paper" work. She thinks she learns better from formulas, but she is not opposed to using the simulations in class, as long as they are accompanied by formal methods.

> I think this [method] is reinforcement of formulas. If I saw both, I think if I had them both [formal and formal methods], I would have understood them better, I would have a better grasp of what exactly it was.

Casey also believed that different types of student would benefit more from different methods, saying "I am an old school, but my sister would prefer this by far because she is good with computers. It just depends on the type of person."

Randomization Test

Exploring the simulation. Casey predicted that the fish oil diet is more effective than the regular oil diet based on the given data. She correctly interpreted the initial graphs, stating "The *x*-axis would be your blood pressure reduction and then the *y*-axis shows whether it's fish or regular." She also understood the re-randomization process. Casey explained, "The people got reshuffled into different groups… The numbers did not change because in order to change the scores, you have to do the experiments again." Although she did not figure out herself why the re-assignment of the participants was

possible into groups, she seemed to understand my explanation, saying, “oh, so if the diets are equally effective, it does not matter which diet they are on.”

Casey also correctly explained the distribution of shuffled mean differences, “So the *x* shows you the differences in means and *y* shows how many times you got those differences in means.” She also was right in saying “according to the graph, [the differences] between -4 and 4 are likely and those in the tails are unlikely.”

While Casey was comfortable with the *recognition*, *integration*, and *comparison* stages, her conclusion was incorrect. She reasoned, “We got -7.714 and the *p-value* of .0012, so 12 out of 1000 shuffled differences. It’s not likely. Which means that they are probably equal to each other.” Casey made a wrong conclusion because she was using a wrong assumption for constructing the empirical distribution. Since she noticed from the original data that the fish oil was better and we were working with the same data to construct the distribution, she thought a small *p-value* rejects the hypothesis that the fish oil is better.

> By this [pointing at the original data], I would conclude that the fish oil is better. By that [pointing at the empirical distribution], I would conclude that they are the same because you are keeping the values, you are just switching between graphs.

When I emphasized that the graph represented possible differences in means for equally effective diets, Casey quickly corrected herself, “Now I understand when you assume they are the same, that’s your null and you want to prove that wrong.”

Student’s view of the simulation. Casey reported liking the randomization simulation more than the binomial program even though she thought the binomial

program looked simpler. She appreciated that the second simulation displayed the original graph together with the sampling distribution – although, it could be the reason she was initially working under a wrong assumption, since the data clearly showed that fish oil blood pressure reductions were greater than those of regular oil.

Casey felt that the program was easy to use and helped her understand the construction of the empirical distribution and the meaning of the *p-value*.

> I think if I saw this [simulation] in class, it would have sculpted my learning in a way that I would have understood the *p-value* a little bit more…proving the null hypothesis wrong because when you do it on pen and paper, you don't see 1000 trials, you just see that one trial.

The program allowed Casey to understand the process of obtaining sampling distribution by repeatedly reassigning participants into groups.

Student's learning preference. Casey still wanted to see both instructional methods in class. In her view, each method has its strength and weaknesses. With simulations, students do not perform computations by hand and therefore might not understand the theory behind the process. However, with formulas, students only see the end result of the graph, and might not know what the empirical distribution and the *p-value* represents.

> I think the blended [approach] would be better just because you get both. If you only have the formula than it's hard to understand how it's done and when you have only this [simulation], you have a computer kind of doing everything for you, so you don't exactly understand the formula. So, I think that the blended method would be better to learn in class.

Derek

Student Background

Derek is pursuing an associate's degree in Education. He is a 21-year-old sophomore who took elementary statistics with instructor Cassidy as a five-week course over the summer three months before the interview and received a B. Derek explained that the course was very fast-paced and did not provide enough time to grasp the material. He is retaking the course with instructor Morgan and feels that he has a better understanding of the concepts. During our interview he only covered the descriptive statistics portion of the course. Therefore, Derek was only exposed to inferential statistics once in his summer course.

Student's Prior Knowledge of Hypothesis Testing

Derek remembered that each hypothesis test involved two hypotheses and that the procedure for testing a claim consisted of five steps. He also recalled, "You never accept the hypothesis." Derek meant that the null hypothesis is never proven, but he did not remember the reason for it. He also recalled that there are two methods of conducting a hypothesis test – critical value and *p-value*. He clarified, "*p-value* only works if alpha was something greater than five, maybe?" It appears that Derek vaguely remembered that when a *p-value* was less than typically 5%, the null hypothesis was rejected. However, he incorrectly thought that the value of alpha determines the method used for hypothesis testing. Clearly, all his recollection dealt with remembering procedures rather than big ideas.

> No, I definitely did not understand them [the concepts]. I understood the process of steps. I also did not understand how to pull the information out of the question and put it into the right places of the steps. I can work through equations all day long, it's ok. What portions I put where, was the most difficult part.

It appears that Derek was confused with symbolic representations and was unsure about the meanings of the notations used in the formulas.

Binomial Simulation

Exploring the simulation. Derek correctly predicted that the distribution for the number of heads out of 10 flips "would have the most number of five's. To have a higher concentration one way or another would be unlikely, and the least likely possibilities would be to get ten heads or ten tails."

Derek concluded that a coin producing nine out of ten heads would be loaded and was able to relate his decision to the empirical distribution, "You would expect this [pointing on 9 on the *x-axis*] to be taller for this graph with your own fair coin." Derek meant that if the coin was fair, the likelihood of flipping nine out of ten heads would be larger. When we calculated the tail proportion of 20/1000, he correctly explained, "If you would have done this [experiment] 1000 times, you would expect that 20 of those times you would get nine or ten heads." Derek realized that the tail proportion is unlikely, and explained, "This is very, very far away from the average of five."

While Derek's conclusion was correct, it seems that he did not understand that nine heads was obtained for one coin using one experiment.

> So we start with saying that we know the fair coin produces for 10 flips an average of five heads. Your set of coins when flipped produced an average of nine, which is far from five.

When I pointed out to Derek that we were only using one coin and flipping it ten times once, he quickly corrected himself and claimed that this was clear to him. However, a couple of other times he used the term "average of nine," which indicates that he was either using an incorrect term or he was thinking that the experiment was done more than once. It is possible that the process of repeatedly tossing ten coins when constructing the empirical distribution made him incorrectly believe that the coin we were testing was also flipped ten times over and over.

Derek found the drug problem more confusing. He correctly set up the simulation and interpreted the distribution for drug X, "So the horizontal axis is showing the number of improvement using drug X out of 1000 samples of 20." Derek then looked at drug Y results and explained, "So if Y is as effective as X, it's unlikely that your sample would have had these results."

While Derek demonstrated an understanding of the sampling distribution and related it to the drug Y result, he had a difficulty making a conclusion with a tail probability. He correctly calculated the tail proportion for 12 improved patients to be 6.2% and concluded, "It's a low percentage, so Y is better." However, he struggled to decide when he would consider results to be non-significant. First, he decided to take 30% as a cut-off proportion. He then found the tail proportion for 11 patients to be 13% and reported:

> I don't know if there is a statistical cut-off, but I suppose we can keep going as long as Y is greater than 50%... no, no, no...[pause] the cut-off is 40% because at this point Y is still better than X… Y still moves your average up.

It seems that Derek lost the link between the number of improvements and the tail proportion. He struggled to explained the meaning of the *p-value*, stating, “There is a 5.6% chance that drug Y is as effective as X.” When I prompted Derek to explain the tail proportion in terms of the number of improved patients, he said, “I think I get it. There is a 5.6% chance of 12 successes occurring, or more.”

Derek correctly concluded that if the number of improved patients was higher than 12, he would have more evidence that “Y is better than X.” However, if this number was lower than 12, he would conclude, “They might be the same.” He was also able to use the tail proportion to make his conclusion, saying, “I see, it [5.6%] is fairly unlikely, but depends on the cut-off,” and understood why they were comparing *p-value* with alpha in class.

In summary, Derek easily moved through the *recognition* and *integration* stages but required guidance at the *comparison* and *explanation* stages.

Making connections between formal inference and SBI. For the coin problem Derek was not able to relate the simulation to its formal method, asking “Is the tail proportion the test statistic?” However, after, the drug problem, he explained “if *p* is low, the null has to go, that what it was! Because it’s such a low probability, Y is not as effective as X, it’s better.” Although Derek did not remember what the null and alternative hypotheses were, he understood that the empirical distribution was constructed under the assumption that the success rate of X is 40%, and why this claim was rejected made sense to him.

Student's view of the simulation. Derek reported liking the program because of its simplicity and the visualization.

> The visual is very, very helpful for me. This is something I was never given before. I like that you very clearly have the places to put each individual piece of information because that's one of the problems I mentioned in the beginning, I did not know what part goes where. Now I know and I also understand the main curve we are using is for X and we are comparing Y to X. I never really got that. We are using two drugs, but we are starting with X, that is our base data and what's the probability that Y is within than data.

Derek appreciated that the program did not require the knowledge of various symbols because their descriptions were written out in words, such as, "Probability of success," "Number of samples," etc. The program also helped him understand that the distribution is based on a certain assumption, which he never realized before. The sampling *z* distribution they used in class was standardized and centered round zero. Therefore, it was unclear that it was constructed for the parameter given in the null hypothesis. The binomial simulation showed clearly that the sampling distribution was for drug X and that the result for Y was compared to it.

Derek suggested that it would be helpful if the program included a cut-off value between likely and unlikely outcomes.

Randomization Test

Exploring the simulation. When we moved to the second simulation, Derek seemed to already know the reasoning process as demonstrated below.

> Derek: The fish oil looks far more effective. You still have some zeros, but the average [reduction] seems much higher

> Researcher: Right, but again, could this difference have occurred by chance, even though the two diets are equally effective?
>
> Derek: Sure, it can, especially because of the small sample size. But what's the probability of that happening?

Derek seemed to have a better sense of the logic used in hypothesis testing. He was also pretty quick to understand the re-randomization process. For the first re-shuffling, he inquired, "Did we change what diet they are on? Have we done a new experiment or have we rearranged our data?" When I explained that we are just rearranging numbers because we want to see what are the possible differences in means assuming equally effective diets, Derek quickly understood the idea, saying, "I see, so we are assuming the two diets are the same so it does not matter."

Derek correctly interpreted the graph of the mean differences. He explained, "The horizontal [axis] is the shuffled differences in the means, and your vertical is the number of times that this difference occurred if they [the diets] are the same." He also moved quickly to the *comparison* stage and observed that the difference of -7.714 was "off the charts." He correctly calculated the tail proportion of about 0.5%. While Derek seemed to be comfortable with the first three stages of the process, he struggled with the *explanation* stage, saying, "Your *p-value* is about half of a percent, which means the null must go, so they both [the two diets] are equally effective."

Derek was using the rule he learned in class, but forgot that the graph represented no differences. It seemed that his previous inferential knowledge prevented him from thinking logically. He still was struggling making his conclusions using tail proportions

and was more comfortable looking at how the observed difference was from the mean. When I asked him to forget about the *p-value* rule and remember the assumption we made when we constructed the graph, he quickly corrected his mistake:

> I see. Assuming that they are equally effective, we would suspect that our data [mean difference of -7.714] would fall somewhere near the mean, which is zero, but it didn't. It is way outside, so therefore, we would that the fish oil is way more effective. Because the sample that we have is an extreme outlier if they [the diets] are the same.

Initially Derek tried to blindly apply the rule, remembering that something had to be rejected. Since the question was whether the fish oil diet was more effective, he decided to reject that statement. However, going back to the logic and remembering what the sampling distribution represented helped him to realize that the extreme result of the experiment would mean that the diets are not equally effective.

Student's view of the simulation. Derek liked "the ability to see our original sample and why the graph is normal." He appreciated that the program calculated the shuffled mean differences because "it would be very hard to do this by hand." Derek liked that the process was fast and noted, "You can get caught up in doing each individual part, so that once you've got your results, you don't know any more what they mean." In Derek's view, manual computations are time consuming and often divert from the main goal of the problem.

Student's learning preference. Derek found both simulations helpful and thought they were easier than the formal methods. He shared, "They [the simulations] are more streamlined and requires me to remember less about the process, but not about how

it works." In other words, because the programs do not require memorization of symbols and steps, Derek finds them simpler. However, he appreciates that the simulations still require understanding of the process. "You still have to understand what is given and what you are looking for, it's just an easier way to do it," he reported. Yet, Derek feels it is important for his future profession to know the theory behind the process. Thus, he prefers learning both formal and informal methods.

> I would still rather do these [the simulation methods] because they are simpler, but the educator in me sees the importance of knowing the formulas. Personally, I find technology easier, but if I really want to learn and know why I am getting what I am getting, I do need to know the formulas to a certain extent.

Ethan

Student Background

Ethan is a 20-year-old freshman majoring in music. He has not taken any college level mathematics courses, but he is currently enrolled in honors Mathematics for Liberal Arts. He also took Geometry and Pre-calculus in high school.

Ethan describes himself as being good in mathematics. He likes numbers because "they make sense" to him. He reported enjoying working with formulas and even considers minoring in mathematics after transferring to a four-year school.

Binomial Simulation

Exploring the simulation. Ethan seemed to understand various parts of the screen but incorrectly interpreted the empirical distribution for the numbers of heads. He said, "This [pointing at the *x*-axis] represents the number of heads and the *y*-axis is the number of tails." We started the simulation over, after which he corrected himself, "The *y*

is the number of times this number [of heads] was flipped." He also explained, "Zero, one, nine, and ten heads are unlikely," and that the other numbers are more frequent.

Ethan reasoned that if a coin produced nine out of ten heads, "it's weighted on one side because there are much more heads than tails." While intuitively he knew that the coin would probably be unfair, he could not use the distribution for a fair coin to explain his conclusion, stating, "This graph does not help me with my decision because we don't know if it's a fair coin." With scaffolding, he was able to reason correctly as shown below:

Researcher: What does this tail proportion mean? [Pointing at 0.9%]

Ethan: There is 0.9% chance of getting nine or ten heads out of a fair coin.

Researcher: So, could a fair coin give you nine heads?

Ethan: It could, but it's unlikely. Oh, so it [the coin] is probably unfair.

Ethan still had a difficulty with nonsignificant results. With eight (tail proportion of 4%) and seven (16%) heads Derek still concluded that the coin was unfair. He reasoned, "16% is a high chance but I am not a risk taker." It appears that he was trying to prove that the coin was fair, and we discussed that that was not the purpose. Still, he struggled to reason with percentages. When I asked Ethan to forget about the tail probability and start thinking in terms of the number of heads, he stated:

> With seven heads, I would probably say it could be fair because it can easily happen [with a fair coin]. With eight heads, I would say that the chance of getting eight heads out of ten flips is so low, so I would say it is unfair.

Ethan also needed a lot of guidance with the drug problem. He correctly reasoned that out of 20 patients he expected eight of them to improve with drug X. He also recognized that "the results are variable" from sample to sample. Moreover, Ethan was able to see from the graph that "numbers closer to eight are likely and [those] very far away are unlikely." However, he struggled to compare drug Y results to X. He wanted to change the probability of success in the simulation to 60% because drug Y helped 60% of the patients. He was trying to generate a sampling distribution for drug Y to see if it would be similar to drug X's distribution. I clarified for Ethan that the success rate for X was known to be 40% because it was tested to a very large number of people. However, drug Y was only tested on 20 patients since it is a new treatment. Ethan then reported, "We cannot conclude anything with such a small sample space" but also agreed that having a larger sample was often not feasible. Once I asked him to recall the coin example and use the same logic, he correctly stated:

> I would say that Y is no better than X because seeing an improvement in twelve patients is likely with X…For 14 patients the probability is low for X, so when Y sees that much improvement that leads you to believe that Y is better.

It is interesting that when I inquired about his cut-off value in terms of the tail proportion, he was confused and said, "I would think 40% improvement is a good number," referring to the percentage of patients improving, rather than to the *p-value*. However, when I inquired for how many improved patients he would be comfortable to conclude that drug Y is better than X, he understood the question and responded, "I would say starting from 13 patients."

Ethan admitted that it was hard for him to reason with tail proportions. Although he understood what they meant, he thought comparing values on the horizontal axis was "more intuitive" than comparing the likelihoods. He explained that his "brain works better this way probably because we read left to right." Ethan meant that the numbers on the horizontal axis increase as we move from left to right, but that the tail proportion increases in the other direction, which he found to be counterintuitive.

In summary, while Ethan eventually was able to understand the functionality and logic of the binomial simulation, he required guidance at every stage and especially struggled at the *comparison* stage of the conceptual framework.

Student's view of the simulation. Ethan thought the program was easy to understand and very visual, but it is the wording he often finds confusing. The program itself was not complicated. However, he was not used to "so much explanation." Ethan also shared that, he is not accustomed to learning mathematics with this type of visual representations, but expressed, "Once you learn how to read graph like that, it is easy to figure it out."

Student's learning preference. Ethan shared that he typically does better with numbers than words and this type of thinking is new to him. He explained,

> I've grown up using formulas, so I have no problem with them, and actually having this much visual aid is fairly new for me. In some ways, it is actually more confusing. It's more intuitive, but it's also more confusing because you can wrap yourself in your own mind when you are trying to figure something out intuitively, whereas with a formula, you can plug and chug and get your answers. But there is a downside to formulas. You have to memorize them and sometimes they are big and annoying. I think visuals are easier for minds to comprehend. However, they would be harder to come up with an answer for, whereas with a

formula, maybe we don't really understand it, but it's easier to come up with the answer.

It was difficult for Ethan to decide which method he preferred because each has its own advantages and disadvantages. While formulas might not make much sense, they are easier to use for the sake of obtaining the answer. However, the simulation required logical reasoning, which is more complicated and might lead to confusion. Ethan thought he would benefit the most by the combination of SBI and formula-based methods.

Randomization Test

Exploring the simulation. Ethan moved much quicker through the fish oil example. He was able to explain correctly the screen representations and even understood rerandomization process fairly quickly, as illustrated below:

Researcher: What do you think just happened?

Ethan: So we re-grouped them [the participants], and I am assuming they took the diets again.

Researcher: They actually did not, and that's the key here. You see how these people went up [pointing on the graph] without changing values. They kept their scores.

Ethan: Right

Researcher: Can you think of why?

Ethan: No, that seems counterintuitive.

Researcher: So remember, our goal is to see what are the differences in means assuming…

Ethan: Equally effective diets. Yes, that makes sense. They are the same, that's why.

Ethan correctly interpreted the empirical distribution of re-grouped mean differences and explained, “Between negative six and six [the mean differences] are fairly likely.” He knew he had to compare the result of the experiment to the distribution. Ethan also correctly computed and explained the tail proportion of 0.02%, saying, “The probability of getting a mean difference of -7.714 is very unlikely if the two diets are the same” and concluded that “the fish oil diet is better.”

Ethan seemed to understand the logic for making conclusions, showing a significant improvement at each stage of the reasoning process. He shared, “I knew that the main focus would be this graph [pointing at the distribution of mean differences] … I knew what the goal was and how to think through it.”

Student’s view of the simulation. Ethan found the randomization test easier to understand than the binomial simulation because “there are more numbers involved.” Perhaps it’s not the numbers that helped him but his familiarity with the inferential reasoning from the previous examples.

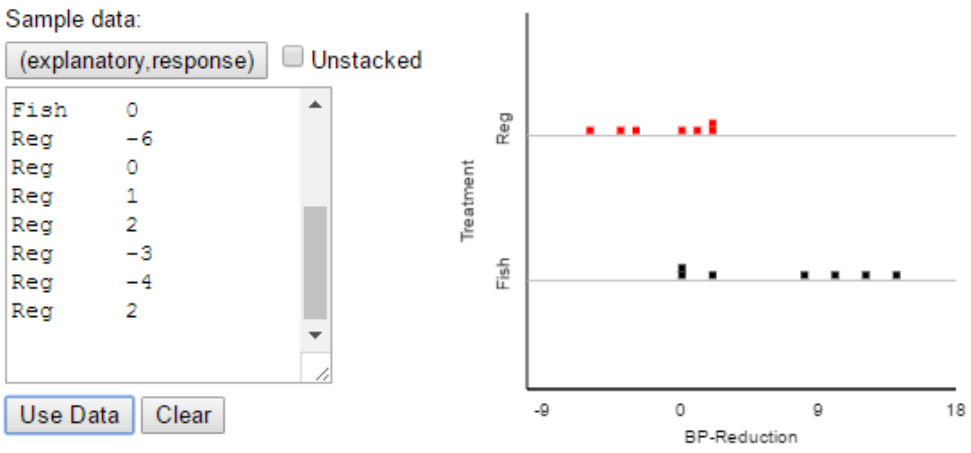
Ethan shared that at first he found the graph of the original data confusing (see Figure 6). He explained,

> I don’t particularly like the way these graphs are set up. I’ve never seen a graph like that before. It almost feels like having two separate graphs combined at once and it’s a little bit confusing.

Ethan initially found the graph difficult because the vertical axis depicted both groups simultaneously – fish oil and regular oil. However, he was able to make sense of the graph without any assistance.

Figure 6

Initial Graph for Randomization Simulation



Student's learning preference. Ethan expressed that because the simulation is very intuitive, it helped him understand statistical ideas. He explained:

> Maybe you won't understand all the ins and outs of how everything works [using the simulation], but you understand the concept better, whereas in algebra you understand the process but you don't understand the outcome. I think this [SBI] taps into different parts of your brain.

Ethan noticed that the program required different skills and understandings than an algebra-based class. He recognized that typically in algebra the emphasis is more on process and less on the meanings of answers. However, with the simulation he had to explain his answers, which required more conceptual understanding. While Ethan thought he would "learn more" from simulations, he believed that both algebraic and visual approaches are useful because they lead to different types of understanding. For this reason, Ethan thought that a blended approach would be optimal for learning inference.

Anne

Student Background

Anne is 23 years old. She used to be in the Liberal Arts program, but recently switched to the Nursing program. She also has been working as a writing tutor for the past two years. Anne has not taken any college level math courses, but completed developmental level classes which are equivalent to pre-Algebra and Algebra I.

Anne shared that she does not "find pleasure in mathematics" and gets easily "frustrated with it." She feels she can do mathematics, but it makes her "feel unhappy," mainly because it does not make sense, and she does not find it relevant to her life.

Binomial Simulation

Exploring the simulation. Even before we started working with the simulation, Anne correctly predicted that for a fair coin out of ten tosses "most likely we get between four and six heads." She was easily able to generate the sampling distribution and interpret its meaning. When I asked her to make a conclusion about the fairness of a coin producing nine out of ten heads, she correctly reasoned, "I would think it would be loaded. I mean, it could be fair because this could just happen one time, but nine out of ten for the first time seems a little bit unlikely."

Anne interpreted the tail proportion of 0.7% as "seven out of 1000 times we are going to get nine or ten out of ten heads for a fair coin, which is very unlikely." Not only did she make the correct interpretation, but she also kept emphasizing that this likelihood

was calculated for a fair coin, making a perfect connection between the result of the experiment and the empirical distribution.

For eight heads Anne concluded, “It’s not as unlikely as with nine heads, but I would still find it unlikely.” However, for seven heads Anne reasoned, “This can easily happen with a fair coin, so it could be fair.”

Despite having a weak mathematical and no statistical background, Anne’s reasoning was flawless at every stage and incorporated uncertainty in her conclusions. She transferred this logic to the drug example. She expected variability in drug X from one group of patients to another as demonstrated below:

> People will be affected differently. There are so many patients with different chemistries. Say, African Americans we treat differently, Caucasians we treat differently, and then Native Americans also differently. There are genetics and so many other factors, like lifestyles…

Anne correctly thought that with drug X most likely out of 20 patients eight would improve. She then constructed a sampling distribution and explained, “The horizontal axis shows the number of successes in 20 patients and the vertical axis shows the number of times it happened.” She then calculated the tail proportion for 12 patients to be 6% and said, “I am inclined to say that Y is probably better than drug X because it’s unlikely that you would get 12 on the first attempt with X.”

Anne related drug Y results with the distribution of drug X. She also recognized that drug Y was tested only once on 20 patients and that this result would be unusual with X. Anne considered 10% as a cutoff between what is reasonable and what is not. She concluded that with 13 improvements (tail proportion of 2%) she “would be more

confident in saying that drug Y is better." However, with 11 patients (14%) she reasoned, "I would not say Y is better. Also, it's a new drug and might be more expensive than X."

Anne went beyond statistical significance and considered practical implications of the results. The combination of her nursing background and her strong written and communication skills might be the reason for her exceptional inferential logic.

Student's view of the simulation. Anne reported liking the binomial simulation because "it's simple, it explains to you each number, and has everything you need." Anne liked that the program provided written explanations for what the input values represented, such as "Number of successes," "Sample size," and "Number of samples." She also appreciated that the simulation showed how the graph was constructed and calculated the probability for different outcomes.

Student's learning preference. When I inquired whether Anne would prefer learning inference through simulations, formula-based method, or with both approaches, she responded:

> I would definitely prefer simulations because I really get bogged down trying to memorize every formula and I also find it annoying, because it's not like we are living in 1800's, and we don't have access to the internet or to books that have formulas in them. I do better problem-solving just to trying to talk it out versus to take that information and put it on to paper and try to figure it out like a math problem with a formula.

Anne also shared that many times in a math class the problems have very unreasonable answers, such as "a person buying 1000 watermelons," which often confused her. She frequently encountered unauthentic math problems that did not make sense, and perhaps that is why she finds math irrelevant. With this program, she did not

see any formulas or symbols and only used logic to reach conclusions. Consequently, Anne reported favoring the use of only simulations for learning inference.

Randomization test

Exploring the simulation. Although Anne predicted that based on the data the fish oil diet appeared to be better, she recognized that "it also depends on people they have" in the two groups. She understood the reshuffling process, "They are just moving the numbers around. I am trying to figure out why you would do this." Once I prompted Anne that we wanted to find the possible differences in means for equally effective diets, she quickly understood the reason, stating, "So assuming they are equally effective, it does not matter which group they [the participants] are in."

Anne explained the distribution of mean differences, "So this axis [pointing at the horizontal axis] is saying what are the shuffled differences and over here [pointing at the vertical axis] is the number of times that you got that difference." Without my help, she knew she had to compare the result of our experiment to this graph and said, "So -7.7 is what we got… so that tells me that fish oil is making a difference because if we assume that it does not matter then it is really unlikely that we are going to get this [result]."

Anne clearly understood the logic of inference and reasoned though the problem with minimal guidance.

Student's view of the simulation. Although Anne thought that the randomization simulation was slightly more difficult that the binomial one, she found it helpful. She explained, "I really like it. I feel it helped me understand the concept more. If we were

talking about it in class, I don't know if I would get it. This [program] really helps me see why the fish diet it better."

Anne also liked that the program calculated means and provided the graph so that she could concentrate on the reasoning process rather than on performing computations.

Student's learning preference. When I shared that some students think that manual computations help them learn, she responded:

> I don't feel formulas help me learn better. I feel formulas frustrate me. I learn by actually thinking about it and I remember how to do it, and why we did it later. I feel if the formulas don't make sense, and you won't remember them later. I feel I usually spend more time memorizing each formula and each symbol and not really focusing on the problem. I think some students like formulas because they are like an anchor for them. It's easier for some people – I just need to plug in numbers, but I don't need to think what they mean.

Anne's argument is reasonable. Instructors Morgan and Scott also shared that many students often find plugging numbers in the formulas easier, but do not want to think what the answer represents. However, according to the GAISE report (2005), the introductory statistics course should develop students' statistical thinking. The inclusion of formulas in an elementary statistics course might hinder the reasoning of students like Anne. She was very happy to hear that some colleges in the United States are implementing simulation-based curriculum and wished that CCCC would adopt it as well.

Student Cross Case Analysis

Below I present a student cross case analysis in order to answer the following research questions:

3) What do community college students who took an introductory statistics course recall and understand about formal hypothesis testing?
4) How do community college students with and without a statistics background understand the functionality and logic of the simulations?
    a. What do students understand and what difficulties do they experience when learning inference through a simulation-based approach?
    b. To what extent do students who had exposure to formal inferential methods make connections between informal and formal inferential methods?
    c. How do students who had exposure to formal inferential methods compare and contrast formal and simulation-based inferential approaches?

Research Question 3. What do community college students who took an introductory statistics course recall and understand about formal hypothesis testing?

The following two assertions were developed by examining the four students' responses:

<u>Assertion 1. Most of these students reported having difficulty understanding hypothesis testing during the course.</u>

Lena, who received an A in the course, was the only student who reported understanding statistical inference in class. The other three students shared having various problems. Natalie and Derek reported not understanding the meaning of the concepts and the reasons behind procedures for hypothesis testing in class. While Natalie was able to follow steps and perform computations, Derek had a hard time remembering symbolic representations and identifying which pieces of information were given. Casey struggled with identifying which test was appropriate to use, as all of them used similar procedures.

Assertion 2. During the interviews, students remembered very little about inference. In addition, most of their retained knowledge was factual rather than conceptual.

None of the students could fully explain the purpose of the hypothesis testing and the meaning of the concepts associated with it. Lena had a partial understanding of inferential ideas. She recalled that hypotheses had to be mutually exclusive and that we examined whether our statistic was reasonable.

All students eventually recalled that the small *p-value* led to rejection of the null hypothesis, but did not remember why and what the null hypothesis represented. Derek also recalled that hypothesis testing could be done two different ways, and that both critical value and *p-value* methods involved five steps. In addition, he remembered that the null hypothesis was never proven.

Natalie recalled statistical terms, such as *p-value* and significance level, but could not remember what they represented. Although Casey reported understanding the concepts in class, she could not explain any of them.

Research Question 4. How do community college students with and without a statistics background understand the functionality and logic of the simulations?

Table 9 depicts a cross-case matrix of students' understanding and difficulties as they move through the four stages of inferential reasoning. I coded student responses on a scale from zero to three based on the amount of guidance they needed to understand the functionality and logic of the simulation. The codes are explained in Table 8 below.

Table 8

Coding Descriptions

Score	**Description**
3	Understood with no guidance
2	Understood with minimal guidance
1	Understood but required a lot of guidance
0	Did not understand

Five assertions were developed after analyzing student responses across cases.

Assertion 1. While exploring the *binomial simulation*, most students had little difficulty at the *recognition* and *integration* stages, but struggled during the *comparison* and *explanation* stages.

Below I describe students' understandings and difficulties during each stage of the conceptual framework.

Table 9

Student Matrix across Four Stages of the Conceptual Framework

		Lena	Natalie	Casey	Derek	Ethan	Anne
Binomial Simulation	**Recognition**	3	3	2	3	2	3
	Integration	3	3	2	3	2	3
	Comparison	2	2	1	1	1	3
	Explanation	3	2	2	2	2	3
Randomization Test	**Recognition**	3	3	3	3	3	3
	Reshuffling	2	0	2	2	2	2
	Integration	3	3	3	3	3	3
	Comparison	3	3	3	3	3	3
	Explanation	2	3	1	2	3	3

***Recognition* stage.** Most of the students could easily interpret the screen representations and sampling distributions, but others had difficulty. Casey initially was unable to distinguish sample size from the number of experiments. She also incorrectly indicated that the horizontal axis of the empirical distribution for the drug example was "the number of trials" and the vertical axis was "the number of patients."

Likewise, Ethan originally struggled with the binomial distribution indicating that the vertical axis represented the number of failures. However, both Casey and Ethan corrected their mistakes after resetting the program and reconstructing the graph.

***Integration* stage.** The *integration* stage was relatively straightforward for most students. While all of them could distinguish likely outcomes from those that are unlikely, Casey and Ethan struggled with tail proportions. Casey considered an outcome to be unlikely if it had less than 50% of occurrence. Ethan was unable to associate the probability of an outcome with the corresponding tail proportion. Initially, it was hard for him to comprehend that as we move from left to right along the horizontal axis, the tail proportion decreases.

***Comparison* stage.** Students found the *comparison* stage most problematic and required guidance to relate the results of the experiment to the empirical distribution.

When asked to make a conclusion about fairness of a coin producing nine out of ten heads, Lena tried to calculate the probability of the coin being fair. However, computing this likelihood is impossible. Likewise, Natalie thought that the distribution of a fair coin was irrelevant because a coin resulting in nine heads was unfair. While Natalie was convinced that that coin was unfair, Ethan reported that the distribution of the fair coin was not useful because we did not know if our coin was fair. Casey also struggled to explain why we use the distribution of drug X to make an inference about the effectiveness of drug Y.

Another problem was comparing the result of the outcome to the mean of the empirical distribution without considering variability in results for the assumed parameter. Three students (Natalie, Casey, and Ethan) reported that drug Y was more effective than X because drug Y improved 60% of the patients, whereas X only helped

40% of people. These students failed to recognize that even though drug X has a true success rate of 40%, it would not produce exactly eight improvements in every sample of 20 patients. Ethan was confused with a similar issue. He did not recognize which success rate represented a sample result and which one was calculated for the larger population. When comparing drug Y to drug X, he wanted to generate a distribution for drug Y, indicating 60% as its true improvement rate.

Half of the students initially struggled with a cut-off value between significant and non-significant results, especially in terms of a tail probability. For Derek, initially the cut-off was 30% or 40%. According to Casey, the fair coin had to produce exactly five heads out of ten tosses. She did not realize that 50% probability is a theoretical value and does not happen for each set of ten tosses. Finally, Ethan also considered a high tail proportion (such as 16%) as a significant result because he was trying to prove the coin was fair.

***Explanation* stage.** There were a couple of problems with student reasoning at this stage, and they are explained below. First, students had difficulty explaining nonsignificant results. Casey would only say that the coin was fair if exactly 50% of the times it landed on a head. Similarly, Derek and Ethan expected a high tail proportion (more than 30%) to conclude that the coin was fair. The problem with their reasoning was that they were trying to prove the fairness of the coin instead of trying to disprove it with extreme results. Derek explained that he thought a tail proportion of 16% was not large enough and thought it was “risky” to conclude that the coin was fair. The finding that

learners struggle more with non-surprising outcomes was found in other studies (Budgett et al., 2013; Holcomb et al., 2010). Fortunately, with a little guidance, all students were able to clear their confusion and correctly interpret results.

Another problem with student reasoning was using deterministic rather than probabilistic language in making conclusions. For instance, Natalie initially deduced, "the coin is definitely unfair," failing to recognize that even a fair coin could produce nine out of ten heads. While Casey and Ethan did not use the term "definitely," they initially concluded that drug Y was more effective than X without considering the chance explanation.

Summary. While all students eventually were able to understand the functionality and logic of the binomial simulation, they initially experienced several difficulties that are summarized in Table 10 below.

Table 10

Student Initial Difficulties with the Binomial Simulation

Recognition	• Confusion between sample size and number of samples • Inability to interpret the *y-axis* of a sampling distribution
Integration	• Inability to relate a tail proportion to the likelihood of the results
Comparison	• Inability to relate results of the experiment to the sampling distribution • Comparing the result to the mean of the sampling distribution, without accounting for variability • Choosing a high cut-off between significant and non-significant results
Explanation	• Attempt to prove the null hypothesis • The use of deterministic language

Assertion 2. While exploring the *randomization simulation*, all students were comfortable at the *recognition*, *integration*, and *comparison* stages. However, some of them struggled to understand the rerandomization process and had difficulty drawing conclusions.

Below I describe students' understandings and difficulties during each stage of the conceptual framework.

***Recognition* stage.** While all students easily understood screen representations and correctly interpreted the empirical distribution, they required scaffolding for the construction of the distribution. For this reason, I represented students' results separately for screen representations and reshuffling as shown in Table 9.

Natalie struggled with the process the most and did not think the participants could be re-assigned into different groups without repeating the experiment. In addition, she did not feel comfortable assuming the equivalence of the two diets because this was the question we were investigating in the first place.

***Integration* stage.** None of the students required guidance at the *integration* stage. Even Casey and Ethan, who experienced difficulties with tail proportions during the binomial simulation were able to reason about probabilities with the randomization tests, showing an improvement in their understanding of the concept.

***Comparison* stage.** All students were able to relate the observed mean difference in blood pressure reduction levels with the empirical distribution. As they explained, they became familiar with the logic from the binomial simulation. It only took the participants couple examples to become comfortable with the reasoning process.

***Explanation* stage.** Although students improved their reasoning with the randomization simulation by incorporating the probabilistic language and considering the chance explanation, other issues immerged.

While Lena realized that the results of the study were very unusual, she was uncomfortable concluding that the fish oil diet was more effective than the regular oil diet. It is interesting that she exhibited a commitment to the null hypothesis for the randomization example, but not the binomial one. This probably happened because the randomization test is more convoluted. During the construction of the sampling distribution Lena kept stressing that the graph is for equal means and reasoned that if the groups had the same effectiveness, the mean difference should fall on that curve. She then concluded that since the difference of -7.714 fell on the graph, the two diets were equally effective.

Casey and Derek recognized that -7.714 was an unusual outcome but incorrectly concluded that the diets were equally effective. Both students remembered from their statistics classes that a low *p-value* disproved a hypothesis. They incorrectly decided to reject the claim that the fish oil was more effective since this was the original research question. However, once I prompted them to again interpret the graph of the differences, Derek immediately arrived at the correct conclusion, whereas Casey needed more guidance. She struggled to understand that the graph was constructed assuming equally effective diets because we used the original data for the construction, which clearly depicted that the fish oil was better. It is interesting to observe that students without prior

statistical knowledge reasoned better than those who have taken statistics. Ethan and Anne did not think about what null and alternative hypotheses represented. They understood that the empirical distribution depicted mean differences for equally successful diets and correctly concluded that an unusual outcome of the experiment disproved the hypothesis of equal effectiveness.

Summary. While the students demonstrated an improvement in reasoning at the *recognition, integration,* and *comparison* stages, three of them initially struggled at the *explanation* stage. In addition, all students required an explanation for the reshuffling process. Student difficulties with the randomization simulation are depicted in Table 11 below.

Table 11

Student Initial Difficulties with the Randomization Simulation

Recognition	• None
Reshuffling	• Inability to understand the logic behind reshuffling in the randomization test
Integration	• None
Comparison	• None
Explanation	• Commitment to the null hypothesis even with low tail proportions • Disproving a wrong hypothesis

Assertion 3: All students expressed mostly positive views of SBI, citing various benefits of the simulations.

Table 12 below summarizes student attitudes toward each simulation.

Table 12

Student Views on Simulations

	Lena	**Natalie**	**Casey**	**Derek**	**Ethan**	**Anne**
Binomial Simulation	• Easy to use • Good visual • Shows construction of the distribution	• Easy to understand • Helpful visual • Helps understand what the *p-value* represented	• confusing because it has "lots of numbers" • hard time differentiating number of samples and sample size	• Simple • Good visual • Does not require memorization of symbols • Clearly shows that the graph is constructed under some assumption	• Easy to use • Helpful visual	• Provides written description of input values • Calculates probabilities • Shows construction of the graph
Randomization Test	• More complex than the Binomial simulation • Needs more guidance • Shows construction of the distribution	• More confusing than the binomial simulation	• Easier than the binomial simulation • Depicts the original data together with the sampling distribution	• Shows why the graph is normal • Performs computations	• Easier than the binomial simulation • The graph for the original data is confusing	• Slightly more difficult than the binomial simulation • Performs computations • Good visual • Helpful for understanding the reasoning process

Binomial Simulation. All students except Casey found the binomial simulation useful for its helpful visuals and simplicity. Natalie reported that the program clearly

explained the meaning of a *p-value*. Lena, Derek, and Anne appreciated that the simulation showed how the empirical distribution was constructed. Anne also liked that the screen not only showed the symbols for input parameters, but explained what each symbol represented. Derek appreciated that the program made it clear that the distribution was constructed based on a certain conjecture (A coin is fair or the two drugs have the same success rate), and we compared the results of our experiment to that hypothesis. Casey was the only student who initially though that the program was confusing and had a hard time understanding what each parameter represented.

Randomization test. Three students (Lena, Natalie, and Anne) thought that the randomization simulation was more difficult than the binomial program compared to two students (Casey and Ethan) who had the opposite view. Nevertheless, all students reported that the simulation was beneficial for various reasons: (1) Shows the construction of the empirical sampling distribution (Lena and Derek), (2) Displays the original data together with the sampling distribution (Lena and Derek), (3) Performs computations (Derek and Anne), (4) Contains good visuals (Anne), and (5) Makes it easy to understand the reasoning process (Anne).

Natalie thought that the program would be easier if it did not show how the graph was constructed. She was confused with the re-assignment of the participants and felt that the process diverted her from the main goal. She though it would be more beneficial if the static graph of the mean differences was presented instead of the dynamic visual of re-grouping.

Derek reported that the graph of the original data was confusing because the vertical axis represented two different diets but was able to read the graph without any assistance.

Assertion 4: Most students reported that formal and informal methods have their own strengths and weaknesses. For this reason, five out of six students expressed preference for a blended learning approach.

Students provided various benefits and drawbacks for each learning approach as illustrated in Table 13. Although Ethan and Anne have not been exposed to formal hypothesis testing, they made a comparison between SBI and formula-based learning in their algebra classes.

All six students thought that simulations make it easier to understand inferential concepts because of their visual aids. However, not all learners are used to this learning method. Ethan expressed that he was used to learning mathematics algebraically and the reasoning through a solution process was new to him. Similarly, Casey reported that students who are not used to working with technology might experience a steep learning curve with simulations. Finally, Natalie shared that some visuals are helpful and others are overwhelming. The binomial simulation helped her understand the concepts, whereas the reshuffling of data in the randomization test confused her.

Table 13

Pros and Cons of SBI and Formal Inferential Methods

	Pros	**Cons**
SBI	• Easy to understand concepts because of the visuals (All students) • Does not require memorization of formulas and symbols (Natalie, Derek, Anne) • Does not require manual computations (Derek, Anne) • Frees up time for reasoning (Anne)	• Too many visuals can be overwhelming (Natalie, Ethan) • Difficult for less computer savvy students (Casey) • Requires sophisticated reasoning (Ethan) • Does not provide theory (Derek)
Formal Inference	• Easy to Come up with the answer (Ethan) • Deepens knowledge by providing the math behind the process (Derek)	• Formulas might be complex and require memorization of symbols (Ethan, Anne) • Formulas often don't make sense (Ethan, Anne) • Formula-based methods hinder learning as the emphasis is on steps and procedures and not on understanding (Anne)

While Derek and Anne appreciated that the simulations performed computations for them, Natalie and Casey reported that manual calculations are needed as they enhance learning. Similarly, Lena expressed that working out problems by hand helps reinforce the concepts learned through visualizations. Finally, Ethan thought that it was important to know theory behind hypothesis testing, which simulations do not provide.

All students except Anne viewed a blended approach as an optimal method for learning statistical inference as illustrated in Table 14. Lena and Casey expressed that depending on their learning preference, different students would have varying benefits from the two methods. While formula-based approach might work the best for mathematically inclined students, simulations would benefit visual learners and those

who are proficient with technology. In addition, the combination of the two methods would help students deepen their understanding of inference.

Similarly, in Ethan's view each method requires a different type of understanding. While formula-based approach stresses knowledge of procedures and the ability to perform computations, SBI promotes conceptual understanding and critical thinking skills. He also pointed out that it is often easier to come up with a right answer with formulas, but the result might not always make sense, while with SBI, you have to have a good grasp of concepts in order to reason correctly. Although it is more difficult to obtain the right conclusion with simulations, they ensure comprehension of the ideas.

Anne was the only student who was in favor of learning only through simulations. She expressed that memorizing symbols and working out formulas prevent her from thinking logically and reasoning about problems. Out of the six students, Anne demonstrated the highest degree of understanding in all stages of the process despite having the lowest mathematics pre-requisites.

Table 14

Students' Learning Preference

	Lena	**Natalie**	**Casey**	**Derek**	**Ethan**	**Anne**
Formal	✓	✓	✓	✓	✓	✗
Binomial	✓	✓	✓	✓	✓	✓
Randomization	✓	✓	✓	✓	✓	✓

Assertion 5: After using the simulations all students who took an introductory statistics course made some connections between SBI and formal methods and understood better the concepts learned in their statistics classes.

All students realized that a tail proportion in the simulations corresponded to a *p-value* in formal methods. While all of them remembered, "when *p* is low the null must go," the program helped them understand why that was the case. As Derek reported, he had no idea prior to using the simulation what a *p-value* represented. Part of it was because he never realized that the sampling distribution was constructed for the null hypothesis. With formal methods, it is often unclear what *z* or *t* distributions refer to, as they are standardized. The simulations depict non-standardized distribution and therefore make it more clear that they are constructed for the value of the parameter stated in the null hypothesis.

Like Derek, Lena also did not remember much about null and alternative hypotheses. However, once she realized that it is used to construct the graph, she was able to state it correctly for every problem.

Casey remembered the least about formal tests, so making any connections was difficult for her. She did realize that a tail proportion was equivalent to a *p-value* and seem to understand why the null was rejected with low tail proportions.

Trying to make connections between formal and informal methods can impede student reasoning. For instance, Natalie found a tail proportion to be small and rejected a wrong claim. She remembered a low *p-value* disproved some hypothesis, but incorrectly

rejected the alternative hypothesis. When I encouraged her to reason based on sampling distribution instead of trying to remember what she learned in class, she made the correct conclusion.

6. RECOMMENDATIONS

Results of the study suggest the following recommendations for teaching inference in an introductory statistics course:

Recommendation 1: A combination of formal and informal inferential methods should be used in an introductory statistics course.

The results of the study revealed several benefits of teaching hypothesis testing through simulations. Instructors and students valued the visual representations of statistical concepts. They appreciated that the programs clearly depicted how empirical distributions were constructed and the assumption they relied on. Another advantage of the simulations is that they use actual data to obtain a sampling distribution instead of relying on mathematically ideal but practically unrealistic normal distributions. Finally, the simulations perform necessary computations allowing users to focus on the interpretation of concepts and reasoning. Incorporating simulation in teaching hypothesis testing directly aligns with GAISE's (ASA, 2005) recommendations, which include utilizing technology, using actual data, and stressing conceptual understanding of statistical ideas.

While the study participants expressed mostly positive views on the simulations and demonstrated an understanding of concepts, they did not support abandoning formal methods for various reasons. Instructor Scott felt that statistics is based on mathematics. Therefore, exposing learners to the mathematical theory behind statistical methods is crucial. Instructor Cassidy shared that the formal methods provide structure, allowing

students to follow specific steps and observe similarity between various statistical tests. Students, on the other hand, felt that formal methods provide a different representation of the material and help reinforce concepts.

Morgan was the only instructor advocating using only simulation. While I agree with her that students who will never need to perform statistical tests only need to understand the main logic of inference, some students taking an introductory statistics course are pursuing careers requiring further statistics knowledge. My concern is that students taking advanced statistical courses will lack required pre-requisite skills and will have difficulty transitioning from SBI to theory-based methods. Currently there are no simulation-based methods for more complex tests such as MANOVA, multiple regression, and others. Therefore, knowledge of symbols, theoretical distributions, and test statistics is necessary for more advanced statistical procedures. For this reason, a curriculum that is solely based on simulations might disadvantage students in need of further statistical training.

Some researchers argue for abandoning the formal methods so that students focus on the logic and not memorization of symbols and steps (Cobb, 2007; Garfield and Ben-Zvi, 2008). The results of the present study showed that students understand the logic fairly quickly. The participants were exposed to two simulations for less than an hour and demonstrated an improvement in reasoning. Incorporating SBI into the curriculum should not take a substantial amount of time and could be achieved by giving up the use of tables and manual computations, which instructors like Cassidy still rely on. Instructors can use

simulations to introduce the logic of inference and core inferential concepts, after which they can switch to theory-based methods, making connections between the two. This way students will be exposed to both methods, empirical and theoretical, and will have an opportunity to understand the meaning of the concepts as well as the mathematics behind them. Finally, by combining traditional and simulation-based methods, the curriculum will no longer be centered on a normal distribution, as criticized by Cobb (2007), and will incorporate randomized data production.

Recommendation 2: A simulation-based approach should be introduced before formal methods. Furthermore, the binomial simulation should precede the randomization test.

Although the purpose of this study was not to determine the details of implementing SBI, the results revealed that students exposed to formal methods did not understand SBI better than those who did not have statistical background. On the contrary, students with no prior inferential knowledge provided a better explanation of the results. Casey and Derek disproved an incorrect statement with a low tail proportion. Trying to recall procedures learned in class hurt their reasoning. On the other hand, Ethan and Anne did not have this problem because they did not know what the null and alternative hypotheses represented and simply relied on logic to make the right conclusion. While results of a small and voluntary sample cannot be generalized to all students, it is natural to introduce a big idea of inference using simulations before going

into details of symbols, formulas, and terms. Instructors Scott and Morgan also recommended working with simulations before introducing formal methods.

Most of the participants of the study found the binomial simulation more intuitive than the randomization test. The binomial simulation is designed for univariate data compared to the randomization test, which involves the comparison of two populations. Therefore, it is reasonable to introduce the binomial simulation before the randomization program.

Recommendation 3: Instructors should be offered a professional development opportunities (e.g., workshops, small group sessions, instructional coaching) for understanding and implementing SBI.

Instructors might need assistance in understanding simulations and using them in their classrooms. Three out of four interviewed instructors were initially puzzled with the re-shuffling in the randomization simulation. While two of them (Morgan and Clark) understood the reasoning process fairly quickly, Cassidy struggled with it even after my explanation. She felt uncomfortable using simulations for teaching inference because this approach was new to her.

Teachers should be not only comfortable with using the simulations, they should become aware of students' common difficulties with them. Although most students eventually were able to reason with simulations, the visuals initially created misunderstanding in some of them (see Tables 10 and 11). After observing the construction of the empirical distribution for the number of heads of a fair coin, Derek

incorrectly concluded that the coin we were testing was also flipped ten times repeatedly. Similarly, Ethan thought that drug Y was testing on many sets of 20 patients. This agrees with Rossman and Chance's (2014) finding that some students incorrectly believe that simulation replicates a research study.

Teachers need to be able to scaffold students through the reasoning process, especially when relating results of a study to an empirical distribution. Five out of six students needed guidance at the *comparison* stage. They either found the outcome from an experiment irrelevant to the sampling distribution or failed to account for variability in the distribution. Finally, students had difficulty reasoning with non-significant results, choosing a high cut-off between significant and non-significant outcomes and attempting to prove the null hypothesis. Instructors should be mindful of students' difficulties in order to foster their understanding of inferential concepts.

Recommendation 4: Continue to explore student understanding and difficulties with SBI.

Although the study showed that students are able to understand inferential reasoning with simulations, students' comprehension of hypothesis testing should be examined further by future studies. Transitioning from SBI to formal methods might create confusion and a cognitive overload. Instructor Scott expressed a concern for using the randomization simulation because conceptually it is different from its corresponding formal test. In addition, Natalie was very confused with regrouping participants in the fish oil example. Even after my explanation, she still did not agree that reshuffling the

participants was justifiable. Like instructor Scott, she thought that while the binomial simulation was helpful for reinforcing concepts, the randomization program introduced a different idea, which she found counterintuitive. It is possible that seeing the simulation before the formal test would be more beneficial, but this possibility requires further investigation.

REFERENCES

American Statistical Association. (2005). *GAISE college report*. Retrieved from http://www.amstat.org/Education/gaise/GAISECollege.html

ARTIST (2006) [website]. Available at https://apps3.cehd.umn.edu/artist/tests/index.html

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentations. Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf

Biau, D. J., Jolles, B. M., & Porcher, R. (2009). P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. *Clinical Orthopaedics and Related Research® Clin Orthop Relat Res, 468*(3), 885-892.

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. (2011). Introductory statistics unconstrained by computability: a new Cobb salad. *Technology Innovations in Statistics Education, 5*(1).

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. V., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests. *Journal of Statistics Education, 17*(2).

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

Cobb, G. (2007). The introductory statistics course: a ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved from: http://escholarship.org/uc/item/6hb3k0nz

Conference Board of Mathematical Sciences. (2010). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States.* Retrieved from http://www.ams.org/profession/data/cbms-survey/cbms2010-Report.pdf

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2006), *Comprehensive Assessment of Outcomes in a First Statistics Course* [Measurement Instrument]. Retrieved from https://apps3.cehd.umn.edu/artist/caos.html.

Erickson, F. (1986). Qualitative methods in Research on teaching. In M. Wittrock (Ed.), Handbook of Research on Teaching, (3rd Ed.), pp. 119-161. New York: Macmillan.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fitch, A., and Regan, M. (2014). Accepting the challenge: constructing a randomization pathway for inference into our traditional introductory course. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_4A1_FITCH.pdf

GAISE (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report.* The American Statistical Association (ASA). Retrieved from http://www.amstat.org/education/gaise/GaiseCollege_full.pdf

Garfield, J. & Ben-Zvi, D. (2007). How students learn statistics revisited: a current review of research on teaching and learning statistics. *International Statistical Review*, 372–396.

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

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: implementing a statistical reasoning learning environment. *Teaching Statistics*, *31*(3), 72-77.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*(7), 883-898.

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*(3), 327-342.

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*(3), 327-342.

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

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_C208_GOULD.pdf

Guba, E.G. & Lincoln, Y.S. (1994). Competing Paradigms in Qualitative Research. In NK Denzin & YS Lincoln (Eds), *Handbook of Qualitative Research* (pp. 105-117). Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

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

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557-560.

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 Teaching Statistics, Ljubljana, Slovenia. Voorberg, The Netherlands: International Statistical Institute.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (∝'s) in classical statistical testing. *The American Statistician*, *57*(3), 171-179.

Knapp H. & FitzGerald, G. (1989). The antihypertensive effects of fish oil: a controlled study of polyunsaturated fatty acid supplements in essential hypertension *JAMA: Journal of the American Medical Association*, *262*(15), 1037-1043.

Konold, C. and Miller, C.D. (2005). *TinkerPlots: Dynamic Data Explorations* [software, Version 1.0]. Emeryville, CA: key Curriculum Press.

Lee, H., Angotti, R., Tarr, J. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the 2003 IASE Satellite Conference on Statistics Education and the Internet*. Berlin. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/6/Lipson.pdf

Liu, Y., & Thompson, P. (2005). Teachers' understanding of hypothesis testing. *In S. Wilson (Ed.), Proceedings of the 27th Annual Meeting of the International Group for the Psychology of Mathematics Education*. Presented at the PME-NA.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The Reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152-173. doi:10.1080/10986065.2011.538301

Makar, K. & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105.

Manor, H., Ben-Zvi, D., & Aridor, K. (2014). Students' reasoning about uncertainty while making informal statistical inferences in an "integrated modeling approach". In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education: Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).*

Marshall, C. & Rossman, G.B. (2010) *Designing Qualitative Research, 5th Edition.* Thousand Oaks, CA: Sage Publications.

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). 1-20.

Maxwell, J.A. (2005). *Qualitative Research Design: An Interactive Approach.* Thousand Oaks, CA: Sage Publications.

Miles, M.B., Huberman, A.M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Chapter 4: Fundamentals of Qualitative Data Analysis. Pp 69-104.

National Council of Teachers of Mathematics. (2000). *Principles and standards for*

school mathematics. Reston, VA: Author.

Nikiforidou, Z., Lekka, A., & Pange, J. (2010). Statistical literacy at university level: the current trends. *Procedia - Social and Behavioral Sciences*, 9, 795-799.

Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, *12*(2), 59-70.

Noll, J. & Hancock, S. (2015). Proper and paradigmatic metonymy as a lens for characterizing student conceptions of distributions and sampling. *Educational*

Paparistodemou, E., Meletiou-Mavrotheris, M. (2008). Developing young students' informal skills in data analysis. *Statistics Education Research Journal*, 7(2), 83-106.

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

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1-2), 27-46.

Pfannkuch, M., Regan, M., Wild, C.J., & Horton, N. (2010). Telling data stories: essential dialogues for comparative reasoning. *Journal of Statistics Education, 18*(1). Retrieved from http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf

Pfannkuch, M., Regan, M., Wild, C., Budgett, S., Forbes, S., Harraway, J., & Parsonage,

R. (2011). Inference and the introductory statistics course. *International Journal of Mathematical Education in Science & Technology*, *42*(7), 903-913.

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal, 7*(2), 107-129.

Rossman, A. (2008). "Reasoning about informal statistical inference: one statistician's view. *Statistics Education Research Journal*, 7(2), 5-19.

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics WIREs Comput Stat, 6*(4), 211-221. doi:10.1002/wics.1302

Rossman, A. J. & Chance, B. L. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *The College Mathematics Journal 30*(4), 297-305.

Rossman, A. J. & Chance, B. L. (2009A). *One proportion inference.* Retrieved from http://www.rossmanchance.com/applets/OneProp/OneProp.htm

Rossman, A. J. & Chance, B. L. (2009B). *Randomization test for quantitative response two means.* Retrieved from http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton: Chapman & Hall.

Stake, R.E. (2006). *Multiple case study analysis.* New York: The Guilford Press.

Tintle, N. (2015). Simulation-based inference in statistics education: Exciting progress and future directions. Retrieved from

http://www.statisticsviews.com/details/feature/7293032/Simulation-based-inference-in-statistics-education-Exciting-progress-and-future-.html

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

Tintle, N., Topliff, K., VanderStoep J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1), 1-24.

Thompson, P.W., Saldanha, L.A., Liu, Y. (2004). *Why statistical inference is hard to understand*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute: Proceedings of the Fifty-Second Session of the International Statistical Institute, 58*, 201-204.

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypothesis testing by university students. *Themes in Education, 3*(2), 183-198.

Wild, C.J., Pfannkuch, M., Regan, M., Horton, N. (2011). Towards more accessible aonceptions of statistical inference. *Journal of the Royal Statistical Society,*

174(2), 247–295.

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: SAGE Publications.

Zieffler, A., Garfiedl, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40-58.

APPENDIX A
Instructor Interview Questions

Before Simulations

1. How long have you taught an introductory statistics course?
2. Describe your general experience with teaching an introductory statistics class.
    a. What statistical topic do you like teaching most and why?
    b. What statistical topic do you find most challenging to teach? Why?
3. Describe your students' experiences with the course.
    a. Do you believe students find the course more or less challenging than other mathematics courses? Why?
    b. In your experience, what are some statistical topics students find generally easy to understand? Why?
    c. What are some areas that student struggle with? Why?

During Simulations

While showing the simulations to the instructors, I will ask them probing questions to assess their understanding of SBI. I will also ask them to conduct a simulation analysis to investigate the following problem:

Suppose that drug X is known to produce improvement in 40% of patients with a particular disease. A researcher wants to show that drug Y is more effective than drug X. He administers drug Y to 20 patients and observes improvement in 12 patients. Can the researcher conclude that drug Y is better than drug X?

After Simulations

1. In general, what do you think about these simulations?
    a. What, if anything, did you like about each of these two simulation-based inferential methods?
    b. Is there anything you find unclear or confusing about each of these two simulations?
2. Do you think a simulation-based method of teaching inference will benefit students or present more challenges for them than traditional methods?
3. Do you think you might use SBI in your classes? If so, how?
    a. Would you use SBI instead of or in addition to formal methods of inference?

Student Interview Questions

Pre-Simulation Questions

1. Which college-level math courses have you taken?
2. Have you taken an introductory statistics course? If yes, when and why?
3. What was your experience?
    a. Which topics did you find easy? Why?
    b. Which topics did you find difficult? Why?
    c. What do you remember about hypothesis testing?

[Make sure the student addresses the purpose of hypothesis testing, the reason for setting up the null and alternative hypotheses, possible conclusions that are made when conducting a hypothesis test, and the meaning of *p*-value]

Guided Exploration of SBI (Binomial Simulation)

1. Given a fair coin, what is the probability of tossing a head?
2. Suppose you tossed a fair coin 10 times. What results do you expect to get?
    a. How many heads and how many tails do you think you are most likely to obtain? Why?
    b. Is it more likely to get 6 heads or 7 heads? Why?
    c. Is it more likely to get 9 heads or 1 head? Why?

[The researcher starts simulating 10 tosses of a fair coin using this simulation http://www.rossmanchance.com/applets/OneProp/OneProp.htm. Then the researcher prompts the student to repeat the experiment several times and to interpret the dot plot of the empirical sampling distribution]

3. If we keep tossing the coin 10 times and recording the number of heads, what do you think the shape of the graph will look like?

[Ask the students to repeat this process until the total number of trials is equal to 100]

4. What percentage of the times did we obtain 5 heads? 9 heads? 10 heads?
5. Now imagine that you have a coin, but you don't know if it's fair. You toss it 10 times and obtain 9 heads. What is your conclusion?
    a. The coin favors heads.
    b. The results occurred by chance even if the coin is fair.
6. What is the likelihood of obtaining 9 or 10 heads if the coin is fair?
7. What does this likelihood tell you about the "fair coin" assumption?
8. Can you summarize the reasoning process you used to reach the conclusion?
9. What if, when we tossed 10 coins, we obtained 8 heads instead of 9? Would your conclusion change? If so, how? If not, why?

[If the student took a statistics course, ask question number 10. Otherwise, present the student with the follow-up task]

10. Do you see any connection between this example and hypothesis testing methods you learned in your statistics class?
    a. What are the hypotheses in this example?
    b. What is an approximate *p-value*?

Follow-up Task

Suppose that drug X is known to produce improvement in 40% of patients with a particular disease.

1. How many improved patients is it possible to get by administering drug X to a sample of 20 patients?
2. How many patients is it likely to get by administering drug X to a sample of 20 patients?
3. A researcher administers drug Y to 20 patients and observes improvement in 12 patients. Can the researcher conclude that drug Y is better than drug X?

Guided Exploration of SBI (Randomization Test)

[Ask the student to read the following]

Researchers investigated whether people on a fish oil diet experienced greater reductions in blood pressure than those on a regular oil diet. They randomly assigned 14 male volunteers with high blood pressure to one of two four-week diets: a fish oil diet and a regular oil diet. Each participant's diastolic blood pressure was measured at the beginning and end of the study, and the change was recorded. The resulting differences in diastolic blood pressure (measured in mmHg) were:

Fish Oil Diet	-8	-12	-10	-14	-2	0	0
Regular Oil Diet	6	0	-1	-2	3	4	-2

1. What do these numbers represent? (What do the numbers -8, 12, 6, 0 mean?)
2. Just by looking at these numbers, which diet do you think is more effective in reducing blood pressure?

[Enter the following data into the randomization applet:

http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2]

Treatment BP-Reduction
Fish -8
Fish -12
Fish -10
Fish -14
Fish -2
Fish 0
Fish 0
Reg 6
Reg 0
Reg -1
Reg -2
Reg 3
Reg 4
Reg -2

3. What is the mean difference between the two groups?
4. Could this difference occur by chance, even if both of these diets are equally effective?
5. If both treatments are equally effective, what possible difference in blood pressure reduction can we get by randomly assigning these 14 participants to two different treatment groups?

   [The researcher asks the student to re-assign the participants into groups]
6. What does the new graph show?
   a. What do the red and black squares represent?
   b. What is the mean difference between groups?

7. If we continue to re-assign participants in groups and to record the mean difference, what do you think the graph will look like?
    a. What is the most likely difference between the means? Why?
8. In our study the mean difference between groups was -7.714. What is the likelihood of obtaining this mean difference given that the two diets are equally effective?
9. Can you conclude that the fish oil diet is more effective in reducing blood pressure than the regular diet?
10. Can you summarize the reasoning process we used to reach the conclusion? [If the student took a statistics course, ask question number 11.]
11. Do you see any connections between this example and hypothesis testing you learned in your statistics class?
    a. What are hypotheses?
    b. What is an approximate *p-value*?

Post-Simulation Questions [will be asked after each simulation]

1. In general, what did you think of this simulation?
2. What, if anything, did you like about the simulation-based method?
3. Is there anything you find unclear or confusing about this method of hypothesis testing?
4. Do you think the simulation methods helped you understand hypothesis testing? If so, how?

    [If the student took an introductory statistics course, ask question 5]
5. Do you prefer the simulation methods or formal hypothesis testing approaches when studying inference, or do you think both methods should be taught in an introductory statistics course? Explain why.

APPENDIX B

Informed Consent Agreement for Instructors

Please read this consent agreement carefully before you decide to participate in the study.

Purpose of the research study: The purpose of the study is to investigate instructors' views for their students' understanding of inferential statistics and instructors' reactions to teaching simulation-based inferential methods.

What you will do in the study: You will be interviewed about your experience with teaching statistics. You will be shown two computer simulations used for performing hypothesis testing informally. You will be asked to report your views about teaching inference using this method. The interview will be audiotaped and transcribed.

Time required: The study will require about one hour of your time.

Risks: There are no anticipated risks to this study.

Benefits: There are no benefits to you for study participation.

Confidentiality: The information that you give in the study will be handled confidentially. To protect your privacy you will be assigned a pseudonym, which will be used during data-collection, analysis, and reporting of findings. After transcribing your interviews, the audio tape will be deleted, and the electronic documents will be stored on a password-protected computer.

Voluntary participation: Your participation in the study is completely voluntary.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty. Should you decide to withdraw, your audiotaped data will be destroyed.

How to withdraw from the study: If you want to withdraw from the study, please tell the researcher to stop the interview. There is no penalty for withdrawing.

Payment: You will receive a $30 gift-card to Amazon.com for participating in the study.

If you have questions about the study, contact:
Irina Timchenko
Telephone: (434)961-5440
Email address: it2h@virginia.edu

Joe Garofalo, Ph.D.
Telephone: (434)924-0845
Email address: garofalo@virginia.edu

If you have questions about your rights in the study, contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: www.virginia.edu/vpr/irb/sbs

Agreement:
I agree to participate in the research study described above.

Signature: _____ Date: _____

You will receive a copy of this form for your records.

Informed Consent Agreement for Students

Please read this consent agreement carefully before you decide to participate in the study.

Purpose of the research study: The purpose of the study is to investigate students' knowledge of statistical inference and their reasoning about and understanding of simulation-based inferential methods.

What you will do in the study: You will be interviewed about your statistical background. After that, you will be shown two computer simulations used for performing hypothesis testing informally. You be asked questions about your understanding of the program components and statistical concepts. You will be then provided with a task and asked how you would use the simulation methods to investigate the given problem.

Time required: The study will require about 1.5 hours of your time.

Risks: There are no anticipated risks to this study.

Confidentiality: The information that you give in the study will be handled confidentially. To protect your privacy you will be assigned a pseudonym, which will be used during data-collection, analysis, and reporting of findings. After transcribing your interviews, the audio tape will be deleted, and the electronic documents will be stored on a password-protected computer.

Voluntary participation: Your participation in the study is completely voluntary.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty. Should you decide to withdraw, your audiotaped data will be destroyed.

How to withdraw from the study: If you want to withdraw from the study, please tell the researcher to stop the interview. There is no penalty for withdrawing.

Payment: You will receive a $20 gift-card to Amazon.com for participating in the study.

Benefits: There is no benefit to you for participating in this study.

If you have questions about the study, contact:
Irina Timchenko
Telephone: (434)961-5440
Email address: it2h@virginia.edu

Joe Garofalo, Ph.D.
Telephone: (434)924-0845
Email address: garofalo@virginia.edu

If you have questions about your rights in the study, contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: www.virginia.edu/vpr/irb/sbs

Agreement:
I agree to participate in the research study described above.

Signature: _____ Date: _____

You will receive a copy of this form for your records.

Student Recruitment Letter

Dear Student,

I am conducting a research study that explores teaching statistical inference using computer simulations. I would like to investigate how students respond to this method of teaching and what their attitudes are toward this instructional approach. I am looking for student volunteers who will participate in the study. Participants will be interviewed about their statistical background. After that, they will be shown computer simulations and will be asked a series of questions, guiding them to solve a task using simulations. At the end of the interview, the participants will be asked about their views on using computer simulations for learning hypothesis testing.

Participation is voluntary and will take up to 1.5 hours of your time. If you chose to participate, you will be provided with a $20 gift certificate to Amazon.com.
If you are interested in being part of the study, please email Irina Timchenko at it2h@virginia.edu.

Sincerely,

Irina Timchenko